RARE: Learn to RAnk and REtrieve for Monocular 3D Object Detection

Hyeonjeong Park[1] Peixi Xiong[2] Xiaoqian Ruan[1] Dian Jia[1] Pei Yu[3] Wei Tang[1]
[1]University of Illinois Chicago [2]Intel [3]Microsoft
{hpark233,xruan9,djia7,tangw}@uic.edu, peixi.xiong@intel.com, pei.yu@microsoft.com

Abstract

Monocular 3D object detection from a single RGB image remains challenging due to two fundamental challenges: the ill-posed nature of 3D localization, where multiple plausible configurations can correspond to the same 2D observation, and unreliable confidence estimation that fails to reflect true localization accuracy. Existing methods predict deterministic 3D boxes that often collapse to implausible mean estimates and rely on absolute confidence scores that are highly sensitive to localization errors. This paper introduces RARE, a unified framework that addresses both challenges through learning to rank and retrieve. RARE formulates confidence estimation as a ranking problem, learning to order detections by their relative quality rather than regressing absolute values. It provides more robust and stable confidence estimates that are less sensitive to localization uncertainty. Building on this improved confidence estimator, RARE learns to construct a query set for each object that predicts multiple diverse and plausible 3D configurations, and retrieves the top-ranked prediction. It explicitly models the multimodal nature of monocular 3D perception and produces more plausible localizations. Extensive experiments demonstrate the effectiveness of RARE. The code is available at https://github.com/HyeonjeongPark37/RARE.

1. Introduction

Monocular 3D object detection aims to detect objects in 3D space from a single RGB image. It is a core perception task for autonomous driving, robotics, and aerial systems, offering a low-cost and easily deployable alternative to the LiDAR setup. Monocular 3D detection involves three subtasks: *localization*, estimating an object's position, size, and orientation in 3D space; *classification*, identifying its semantic category; and *confidence estimation*, assessing how likely a detection corresponds to a true positive. Among these, classification has been largely solved: for example, modern 2D detectors [43, 67] achieve over 90% mAP for car detection on the KITTI benchmark. However, accurate 3D localization and reliable confidence estimation remain significant challenges.

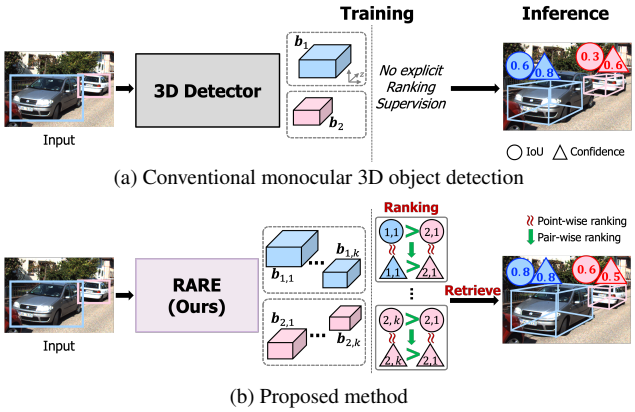


Figure 1. **Comparison between conventional monocular 3D detection and our proposed method.** (a) Conventional detectors predict deterministic 3D boxes that often collapse to implausible mean estimates and rely on absolute confidence scores that are sensitive to localization errors. (b) RARE improves both 3D localization and confidence estimation by constructing a query set for each object that predicts multiple diverse and plausible 3D configurations, and formulates 3D detection as a ranking-and-retrieval process.

To improve 3D localization accuracy, extensive research [16, 30, 35, 43] has focused on developing neural architectures tailored for monocular 3D detection. More recent works [20, 34, 45] integrate geometric modeling with data-driven learning to enhance spatial reasoning. Nevertheless, localization remains difficult because the task is intrinsically ill-posed: multiple plausible 3D configurations can correspond to the same 2D observation, even when geometric constraints are applied (see Sec. 3.2 for a formal analysis). Existing methods predict a single deterministic 3D box for each object, which often collapses into a mean estimate that does not correspond to any plausible solution.

Confidence estimation is equally critical, as it not only directly impacts detection performance, e.g., true positive rate and average precision, but also determines how trustworthy a perception system is. Early approaches [55] sim-

ply reuse the classification score as a proxy for confidence, which fails to reflect the reliability of 3D geometric estimates. Later works introduce confidence measures based on depth uncertainty [34, 67] or 3D box quality [50, 51]. However, recent diagnostic studies [36, 51] as well as our own results (Sec. 3.1) show that confidence estimates still substantially misalign with actual localization accuracy. We conjecture a possible reason is that existing methods regress absolute confidence values, which are highly unstable: small errors in estimated depth or orientation, which are common in monocular setups, can drastically alter the true confidence value.

This paper introduces RARE, a unified framework that enhances both 3D localization and confidence estimation through *learning to rank and retrieve*. RARE is built on two key insights.

Confidence estimation should be relative, not absolute. Instead of regressing only absolute confidence scores, RARE learns to rank detections by their relative quality. This is achieved through an integrated point-wise and pair-wise ranking loss, which enforces both global calibration and local ordering consistency. It eliminates the need to predict the exact confidence values and instead focuses on a more stable supervision signal: comparison between detections within an image. Because the relative ranking of detections is much less sensitive to localization errors, the model learns a smoother, more generalizable mapping from image features to detection reliability.

3D localization should be multimodal, not deterministic. Instead of directly regressing a single 3D box, which tends to collapse to an implausible mean solution, RARE learns to construct a query set for each object that predict multiple diverse and plausible 3D boxes. Each hypothesis represents a distinct possible spatial configuration of the object. From each query set, RARE then retrieves the top-ranked prediction as the final detection, using its learned confidence scores as the retrieval criterion. This process explicitly models the multimodal nature of monocular 3D perception and yields more plausible localizations.

Our main contributions are summarized as follows:

- We formulate confidence estimation in monocular 3D object detection as a learning-to-rank problem. Unlike conventional absolute confidence learning, our joint point-wise and pair-wise formulation produces confidence estimates that are more robust to localization uncertainty.
- Building on the improved confidence estimator, we construct a query set for each object which is learned to predict diverse and plausible 3D configurations and retrieve the top-ranked one as the final prediction. It explicitly models the multimodal nature of monocular 3D detection to produce more plausible localizations.
- RARE integrates the ranking and retrieval formulations into a single detection transformer model trained end-to-end. Extensive experiments demonstrate that RARE outperforms state-of-the-art monocular 3D detectors. We will make the code publicly available.

2. Related work

Monocular 3D object detection lifts 2D detections to 3D from a single image but faces severe depth ambiguity. Many methods mitigate this by adding auxiliary signals, such as LiDAR-based training/distillation [15, 30, 39, 42, 46, 60], CAD priors [24, 33], video [4, 56], or depth [40, 54, 57]. While effective, they require extra sensors or computational overhead, motivating image-only approaches that train and infer from a single image for practical deployment.

Most monocular 3D detectors build on convolutional (conv.) architectures [3, 10, 20, 26, 32, 34, 35, 58, 61], and transformer/DETR-style models have also been explored. For instance, MonoDTR [15] uses a depth-aware transformer, and MonoDETR [67] derives object-level depth labels to improve a depth-aware decoder. MonoDGP [43] models geometry error with a decoupled decoder. Notably, most still retain conv. backbones, which provide multi-scale features and reduce feature size for computational efficiency. Complementary work [29, 37] explores test-time and domain adaptation to address distribution shift. Despite these advances, most image-only methods output a single 3D estimate per 2D detection, which can collapse to an implausible mean across modes. This motivates predicting multiple 3D box candidates per 2D object and raises the central question of how to score and select them effectively.

Set-based prediction is a natural fit for monocular 3D detection, where the mapping from a single image to a 3D box is underdetermined. In 2D detection, DETR-style variants [8, 68, 71] densify supervision with one-to-many label assignment, allowing multiple positives to match a single ground truth, thereby improving recall. For other 3D vision tasks, such as human/hand pose estimation and reconstruction, mixture density networks [1, 21, 64] and multi-head predictors [18, 22] are used to preserve multiple modes. U-CMR [13] proposes a camera-multiplex objective to handle multiple viewpoints by optimizing a soft average of reconstruction errors for single-view reconstruction. However, monocular 3D object detection has been less explored from this perspective. Related ideas exist in anchor-based work [3], which outputs several boxes per object. However, they depend on predefined anchors, making performance sensitive to anchor coverage and class bias. Multi-depth approaches [25, 61, 67] focus on depth cues and can underrepresent overall 3D quality, while introducing additional hyperparameters such as thresholds for depth/score fusion. In contrast, inspired by camera-multiplex [13], we generate a compact, data-dependent set of whole 3D box candidates per object and retrieve the best one with a learned confidence, without fixed anchors or external depth cues.

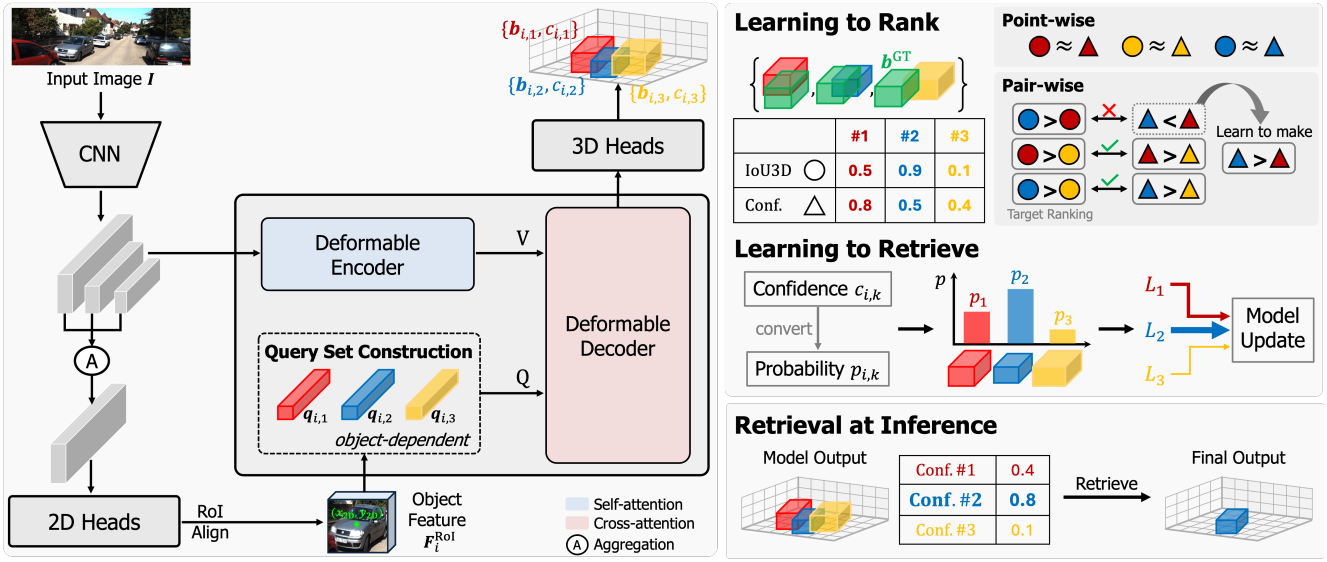


Figure 2. **Overview of the RARE framework.** RARE introduces two new mechanisms for monocular 3D object detection to improve confidence estimation and 3D localization. First, RARE learns to rank detections based on their relative quality rather than regressing absolute scores, which enforces both global calibration and local ordering consistency (Sec. 3.1). Second, RARE learns to construct a query set for each object that predicts multiple diverse and plausible 3D configurations and retrieves the top-ranked one based on the learned confidence ordering (Sec. 3.2). The two mechanisms are integrated within a detection transformer architecture that is trained end-to-end (Sec. 3.3).

Confidence is critical because, in evaluation and downstream decision making, predictions are ordered by the score and then selected according to that ranking, so the score should reflect localization quality (*i.e.*, 3D IoU). In monocular 3D, proxy scores have been proposed: MonoDIS [50] derives a confidence from the 3D box regression loss, and PL [51] encodes order only in the target for relative confidence supervision, updating a single confidence value. Uncertainty-based proxies also appear; for example, GUPNet [34] converts depth uncertainty into a 3D confidence, yet depth alone may not capture full box quality. In the point-cloud regime [47, 48, 53, 62], detectors often optimize a 3D IoU head to guide NMS and refinement, which is effective but not designed to supervise ranking. Meanwhile in 2D, ranking-oriented methods have been extensively explored. They can be categorized into point-wise [17, 23, 44, 66], pair-wise [5, 31, 38], and list-wise [7, 59, 63] losses, depending on whether they supervise individual boxes, box pairs, or entire ranked lists. Point-wise losses push predicted IoUs toward one and use IoU-aware classification confidences for individual boxes; pair-wise losses impose relative ordering between box pairs (e.g., within the positive set or between positives and negatives), and list-wise losses directly optimize surrogates of ranking metrics (e.g., AP) over the entire ranked list. By contrast, we supervise the ranking with a joint point-wise (IoU-aligned confidence regression) and pair-wise (IoU-ordered constraints) objective, without defining positive or negative sets. Rather than pushing IoU predictions toward one, RARE learns a confidence aligned with true 3D box quality so that the score encodes the relative quality of competing predictions and enables reliable retrieval.

3. Method

Monocular 3D object detection aims to infer object classes, 3D bounding boxes, and confidence scores from a single RGB image. We propose RARE, a novel rank-and-retrieve framework, to improve the confidence estimation and 3D localization accuracy. It includes two new mechanisms for monocular 3D object detection. First, RARE learns to rank detections based on their relative quality rather than regressing absolute scores, which enforces both global calibration and local ordering consistency. Second, RARE learns to construct a query set for each object that predicts multiple diverse and plausible 3D configurations and retrieves the top-ranked one based on the learned confidence ordering. We describe the learning-to-rank and learning-to-retrieve formulations in Sec. 3.1 and Sec. 3.2, respectively, and summarize the RARE architecture in Sec. 3.3.

3.1. Learning to Rank

Motivation. Confidence estimation is crucial for monocular 3D object detection, as it affects both detection performance and the trustworthiness of a perception system. Existing methods learn an absolute confidence score for each 3D detection, typically using depth uncertainty or IoU with

the ground truth. However, recent diagnostic studies and our own observations in Fig. 3 show that confidence estimates still substantially misalign with actual localization accuracy. We conjecture that a possible reason is that absolute confidence values are highly unstable: small errors in estimated depth or orientation, which are common in monocular setups, can drastically alter the true confidence. This motivates us to explore a different confidence model that learns to rank detections by their relative quality instead. As shown in Fig. 3, this substantially improves the confidence estimates and detection performance of a MonoDETR baseline.

Ranking-based Confidence Learning. For a detected object with 3D box $\boldsymbol{b}_i$ and class y_i, its *ground truth confidence* $\hat{c}_i$ should depend on both the accuracy of the 3D box and the correctness of the class. Therefore, we formulate the following definition:

$$\hat{c}_i = \max_j \delta(y_i = \hat{y}_j)\texttt{IoU3D}(\boldsymbol{b}_i, \hat{\boldsymbol{b}}_j) \tag{1}$$

where $\hat{y}_j$ and $\hat{\boldsymbol{b}}_j$ are the class and 3D box of the j-th ground truth object, respectively, δ is an indicator function returning 1 for a true statement and 0 otherwise, and `IoU3D` calculates the IoU between two 3D boxes.

RARE learns to rank 3D detections by their relative quality, based on an integrated point-wise and pair-wise ranking loss:

$$\mathcal{L}^{\text{rank}} = \ell^{\text{point}} + \ell^{\text{pair}} \tag{2}$$

The point-wise loss promotes consistency between the confidence estimate and its corresponding ground truth:

$$\ell^{\text{point}} = \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}}(c_i - \hat{c}_i)^2 \tag{3}$$

where $\mathcal{D}$ is the set of 3D object detections predicted from an image, and $c_i \in (0, 1)$ is the estimated confidence. The point-wise loss ensures that the learned confidence predictions are aligned with its true quality. However, relying solely on point-wise supervision makes the confidence estimator sensitive to small localization errors.

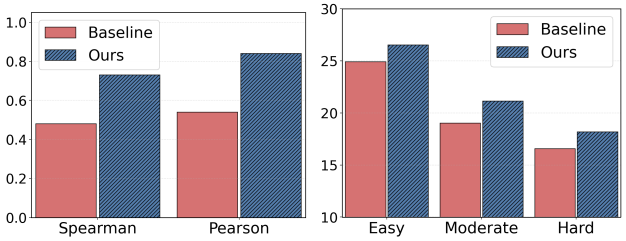
The pair-wise loss enforces relative ordering between pairs of detections. We first define the pair preference label as:

$$\hat{r}_{i,j} = \texttt{sign}(\hat{c}_i - \hat{c}_j) \in \{-1, +1\} \tag{4}$$

where $i, j \in \mathcal{D}$ index a pair of 3D object detections with $\hat{c}_i \neq \hat{c}_j$. Then, the pair-wise loss is formulated as a logistic loss:

$$\ell^{\text{pair}} = \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}} \log(1 + \exp(-\hat{r}_{i,j}\,(z_i - z_j))) \tag{5}$$

where $\mathcal{P}$ is a set of randomly sampled detection pairs, and $z_i \in \mathbb{R}$ is the unnormalized confidence logit that converts to $c_i \in (0, 1)$ through a Sigmoid function. This term explicitly enforces that detections with higher quality receive higher confidence scores, thereby improving discriminability and ranking consistency across examples.



(a) Confidence qualities (b) 3D detection accuracy

Figure 3. **Confidence quality and 3D detection accuracy on KITTI/Car (validation).** (a) Confidence quality evaluated by the rank (Spearman) and linear (Pearson) correlations between estimated and ground truth confidence values. (b) AP_{3D} for all difficulties.

Discussion. Altogether, the point-wise term provides a global confidence calibration, while the pair-wise term refines local ordering relationships among detections. Empirically, we find that using only the pair-wise loss leads to slow convergence and unexplainable confidence scores. The joint optimization of point-wise and pair-wise objectives provides the best of both worlds: the point-wise term anchors the confidence range, while the pair-wise term sharpens the relative ranking.

It is worth noting that our learning-to-rank formulation differs from conventional ranking or sorting losses in 2D object detection [31, 38], which partition predictions into positive (*i.e.*, foreground) and negative (*i.e.*, background) sets and optimize their separation. Such binary separation schemes need to pre-define an evaluation metric-dependent IoU threshold to distinguish positives from negatives. Moreover, these formulations treat all positive or negative samples equally and ignore their relative quality. As a consequence, the model is not explicitly encouraged to rank higher-quality detections above lower-quality ones.

3.2. Learning to Retrieve

Motivation. In monocular 3D object detection, multiple plausible 3D configurations can correspond to the same 2D observation even after applying geometric constraints, e.g., an object has a known physical size and lies on the ground. Inspired by prior work [25, 35, 41, 49, 69], we provide a formal analysis of this phenomenon. Consider a simple yet realistic camera setting, where the optical axis is parallel to the ground. Based on the pinhole camera model, the depth D of an object on the ground can be estimated from the focal length f, the observed object height h in the image, and the known physical height H: $D = f \cdot H/h$. If the visually observed height increases by one pixel, the estimated depth

becomes $D' = f \cdot H/(h+1)$, yielding the depth difference:

$$D - D' = \frac{f \cdot H}{h} - \frac{f \cdot H}{h+1} = \frac{D^2}{f \cdot H + D}. \quad (6)$$

This formulation shows that the depth variation induced by a one-pixel change increases with distance. Due to image quantization, as D grows, a single-pixel difference in image space leads to a substantial change in 3D depth, introducing inherent ambiguity. To ground this observation, consider KITTI with focal length $f \approx 721.54$ pixels and average car height $H \approx 1.5$ m. At depths of 20 m, 30 m, and 40 m, the corresponding depth variations induced by a one-pixel change are approximately 0.36 m, 0.81 m, and 1.43 m, respectively.

As a result, multiple distinct 3D configurations can produce nearly identical 2D observations. A standard single-point regression model therefore tends to predict their conditional mean, often leading to physically implausible estimates. This "mean-collapsing" effect is well known in the machine learning literature [1, 2, 52].

Query Set Construction, Learning, and Retrieval. To address this issue, RARE learns to construct a *query set* for each object that predicts multiple diverse and plausible 3D configurations to capture the multimodal nature of monocular 3D perception. From each query set, RARE then retrieves the top-ranked prediction as the final detection, using its learned confidence scores as the retrieval criterion.

Following Deformable DETR [70], we first locate regions of interest (RoIs) on the feature map obtained by the transformer encoder and then use RoI features to generate content-dependent object queries, which speed up convergence compared to random initialization. Instead of generating a single query per RoI, we construct a set of K queries for each object:

$$\{\boldsymbol{q}_{i,1}, \ldots, \boldsymbol{q}_{i,K}\} = \texttt{MLP}(\boldsymbol{F}_i^{\text{RoI}}),\ i = 1, \ldots, N \quad (7)$$

where $\boldsymbol{F}_i^{\text{RoI}}$ denotes the features of the i-th RoI, and N is the number of RoIs. All queries are processed by the transformer decoder. Each updated query predicts a candidate 3D box $\boldsymbol{b}_{i,k} = (\boldsymbol{x}_{i,k}, d_{i,k}, \boldsymbol{s}_{i,k}, \theta_{i,k})$ along with its confidence score $c_{i,k}$, where $\boldsymbol{x}_{i,k}$ denotes the 2D center, $d_{i,k}$ the depth, $\boldsymbol{s}_{i,k}$ the 3D size, and $\theta_{i,k}$ the yaw. Each set $\{\boldsymbol{b}_{i,1}, \ldots, \boldsymbol{b}_{i,K}\}$ represents K possible 3D configurations for the object in the i-th RoI.

We design a 3D box multi-hypothesis loss to encourage the query set to predict 3D boxes that are both plausible and diverse. Let M be the number of RoIs in a mini-batch. For each RoI, we convert the confidence logits of K hypotheses into a probability distribution: $\{p_{i,1}, \ldots, p_{i,K}\} = \texttt{Softmax}(\{z_{i,1}, \ldots, z_{i,K}\})$. The loss is defined as

$$\mathcal{L}^{\text{3D}} = \frac{1}{M} \sum_{i=1}^{M} \sum_{k=1}^{K} p_{i,k} \ell^{\text{box}}(\boldsymbol{b}_{i,k}, \hat{\boldsymbol{b}}_i) + \sum_{k=1}^{K} |\bar{p}_k - \tfrac{1}{K}| \quad (8)$$

where ℓ^{box} is a common 3D box regression loss (see supplementary material), and $\bar{p}_k = \frac{1}{M} \sum_{i=1}^{M} p_{i,k}$ is the average selection probability of the k-th hypothesis in the batch.

The first term softly aggregates box regression errors across hypotheses, weighting each by its selection probability. This probabilistic weighting encourages the model to pull confident hypotheses closer to the ground truth while providing gentle guidance to less certain ones. The second term regularizes the selection distribution by penalizing deviation from a uniform prior, thereby preventing the model from collapsing its probability mass onto a few dominant hypotheses and promoting diversity within each query set.

During inference, the final 3D detection for each object is retrieved as the top-ranked prediction from the corresponding query set:

$$\{(\boldsymbol{b}_{i,k^*}, y_i, c_{i,k^*}) : k^* = \arg\max_k c_{i,k}; i = 1, \ldots, N\} \quad (9)$$

where the object class y_i is predicted from the RoI features.

3.3. Summary of RARE

RARE integrates learning-to-rank and learning-to-retrieve within a detection transformer architecture that is trained end-to-end. Given an RGB image, a DLA34 backbone [65] extracts multi-scale feature maps, which are then processed by a multi-scale deformable self-attention encoder. Meanwhile, an aggregated feature map is fed to a 2D head to locate RoIs by predicting centerness, 2D size, and 2D offset heatmaps. We retain the top-N RoIs based on their centerness scores. For each RoI, we use RoI Align to extract its features, generate a set of K queries through an MLP, and predict the object class probabilities. A deformable transformer decoder updates the queries by attending to visual tokens obtained from the encoder. Each updated query predicts a 3D bounding box and a confidence score.

The learning objective includes three losses:

$$\mathcal{L}^{\text{all}} = \lambda^{\text{2D}} \mathcal{L}^{\text{2D}} + \lambda^{\text{3D}} \mathcal{L}^{\text{3D}} + \lambda^{\text{rank}} \mathcal{L}^{\text{rank}} \quad (10)$$

where λ denotes a balancing hyper-parameter. $\mathcal{L}^{\text{2D}}$ is mean square error loss between the predicted 2D heatmaps and the corresponding ground truth. $\mathcal{L}^{\text{3D}}$ and $\mathcal{L}^{\text{rank}}$ have been introduced in Sec. 3.2 and Sec. 3.1. At inference, RARE retrieves the top-scoring candidate per RoI as the final prediction based on the learned confidence score.

4. Experiments

4.1. Setup

Datasets and Evaluation Metrics. We evaluate our method on KITTI [12] and nuScenes [6] benchmarks.

- ***KITTI*** includes 7,481 training and 7,518 test images. Following the common split [9], we use 3,712 images for training and 3,769 for validation. The dataset defines

Methods	$AP_{3D\|40}$			$AP_{BEV\|40}$		
	Easy	Mod.	Hard	Easy	Mod.	Hard
CaDDN† [46] ('21)	19.17	13.41	11.46	27.94	18.91	17.19
MonoDTR† [15] ('22)	21.99	15.39	12.73	28.59	20.38	17.14
DID-M3D† [40] ('22)	24.40	16.29	13.75	32.95	22.76	19.83
CMKD† [14] ('22)	25.09	16.99	15.30	33.69	23.10	20.67
LPCG† [39] ('22)	25.56	17.80	15.38	35.96	24.81	21.86
MonoNeRD† [60] ('23)	22.75	17.13	15.63	31.13	23.46	20.97
OM3D† [42] ('24)	25.55	17.02	14.79	35.38	24.18	21.37
MonoTAKD† [30] ('25)	27.91	19.43	16.51	38.75	27.76	24.14
MonoDLE [35] ('21)	17.23	12.26	10.29	24.79	18.89	16.00
GUPNet [34] ('21)	20.11	14.20	11.77	-	-	-
DEVIANT [20] ('22)	21.88	14.46	11.89	29.65	20.44	17.43
MonoCon [32] ('22)	22.50	16.46	13.95	31.12	22.10	19.00
GeoAug [28] ('22)	23.41	15.26	12.80	31.58	20.75	17.66
MonoJSG [27] ('22)	24.69	16.14	13.64	32.59	21.26	18.18
MonoDDE [25] ('22)	24.93	17.14	15.10	33.58	23.46	20.37
MonoDETR [67] ('23)	25.00	16.47	13.58	33.60	22.11	18.60
DDML [10] ('23)	23.31	16.36	13.73	-	-	-
MonoUNI [16] ('23)	24.75	16.73	13.49	-	-	-
FD3D [58] ('24)	25.38	17.12	14.50	34.20	23.72	20.76
MonoCD [61] ('24)	25.53	16.59	14.53	33.41	22.81	19.57
MonoDGP [43] ('25)	26.35	18.72	15.97	35.24	25.23	22.02
RARE	**28.83**	**19.57**	**17.38**	**38.46**	**26.37**	**23.46**

Table 1. **Comparison on *Car* category of the KITTI test set.** All methods follow the official evaluation protocol [12]. Methods marked with † use LiDAR as an auxiliary training source. Best and second-best results are shown in bold and underlined, respectively.

three difficulty levels (Easy, Moderate, Hard) based on occlusion, truncation, and the minimum height of a 2D bounding box. We report $AP_{3D|40}$ and $AP_{BEV|40}$ (bird's-eye view) using IoU thresholds of 0.7 for 'Car' and 0.5 for 'Pedestrian' and 'Cyclist' [50].

- ***nuScenes*** consists of 28,130 images and 6,019 validation images captured from the front camera. We use the validation set for cross-dataset evaluation following [20, 26]. After matching predictions to the ground truths based on an IoU_{2D} overlap threshold of 0.7, we compute the mean absolute error (MAE) of the depths of the predictions and the ground truth boxes [49].

Implementation Details. RARE is trained on two NVIDIA A100 GPUs with a total batch size of 32. We adopt Hierarchical Task Learning (HTL) [34] with a linear warm-up strategy for stable training. To mitigate the limited dataset size, we apply MixUp3D [26] and DivAlign [11] as data augmentation. We use the Adam optimizer [19] with an initial learning rate of 0.001 and train for 800 epochs. We set λ^{2D} and λ^{3D} to 1, while λ^{rank} is set to 10 for ℓ^{point} and 0.5 for ℓ^{pair}. Each RoI is pooled to 7×7 and the number of candidate queries K is set to 3. The transformer uses 3 encoder and 3 decoder layers with 8 attention heads; all hidden dimensions are 256. At inference, detections with 2D scores below 0.2 are discarded, and we do not apply NMS.

4.2. Main Results

Results on *Car* Category of KITTI Test Set. Tab. 1 compares RARE with recent state-of-the-art methods on the KITTI test set for the *Car* category. RARE achieves the highest $AP_{3D|40}$ and $AP_{BEV|40}$ across all difficulty levels, yielding substantial margins over the second-best monocular approach, MonoDGP. Specifically, RARE attains relative gains of **9.4%/4.5%/8.8%** in $\mathbf{AP_{3D|40}}$ and **9.1%/4.5%/6.5%** in $\mathbf{AP_{BEV|40}}$ on the Easy/Moderate/Hard levels, respectively. Despite relying solely on monocular RGB supervision, RARE also remains competitive with LiDAR-augmented approaches such as MonoTAKD†, outperforming it on $\mathbf{AP_{3D|40}}$ with relative gains of **3.3%/0.7%/5.3%** while achieving comparable $\mathbf{AP_{BEV|40}}$ across all difficulty levels. This demonstrates that our ranking-and-retrieval framework is significantly effective at handling the inherent 3D ambiguity in a purely monocular setting.

Methods	Ped., $AP_{3D\|40}$			Cyc., $AP_{3D\|40}$		
	Easy	Mod.	Hard	Easy	Mod.	Hard
CaDDN† [46] ('21)	12.87	8.14	6.76	7.00	3.41	3.30
MonoDTR† [15] ('22)	15.33	10.18	8.61	5.05	3.27	3.19
CMKD† [14] ('22)	17.79	11.69	10.09	9.60	5.24	4.50
OM3D† [42] ('24)	14.68	9.15	7.80	7.37	3.56	2.84
MonoTAKD† [30] ('25)	16.15	10.41	9.68	13.54	7.23	6.86
MonoDLE [35] ('21)	9.64	6.55	5.44	4.59	2.66	2.45
GUPNet [34] ('21)	14.72	9.53	7.87	4.18	2.65	2.09
DEVIANT [20] ('22)	13.43	8.65	7.69	5.05	3.13	2.59
MonoCon [32] ('22)	13.10	8.41	6.94	2.80	1.92	1.55
MonoJSG [27] ('22)	11.02	7.49	6.41	5.45	3.21	2.57
MonoDDE [25] ('22)	11.13	7.32	6.67	5.94	3.78	3.33
MonoDETR [67] ('23)	12.65	7.19	6.72	5.12	2.74	2.02
DDML [10] ('23)	14.90	10.28	8.70	5.38	2.89	2.83
MonoUNI [16] ('23)	**15.78**	10.34	8.74	7.34	4.28	3.78
MonoDGP [43] ('25)	15.04	9.89	8.38	5.28	2.82	2.65
RARE	14.85	**10.79**	**9.04**	**11.17**	**5.96**	**5.28**

Table 2. **Comparison on *Pedestrian* and *Cyclist* categories of the KITTI test set.** All methods follow the official evaluation protocol [12]. Methods marked with † use LiDAR as auxiliary training data. Best and second-best results are shown in bold and underlined, respectively.

Methods	KITTI Val.				nuScenes frontal Val.			
	0-20	20-40	40+	All	0-20	20-40	40+	All
MonoRCNN [49]	0.46	1.27	2.59	1.14	0.94	2.84	8.65	2.39
GUPNet [34]	0.45	1.10	1.85	0.89	0.82	1.70	6.20	1.45
DEVIANT [20]	0.40	1.09	1.80	0.87	0.76	1.60	4.50	1.26
MonoCon [32]	0.40	1.08	1.78	0.85	0.78	1.65	6.02	1.40
MonoUNI [16]	0.38	**0.92**	1.79	0.87	0.72	1.79	4.98	1.43
MonoCD [61]	0.37	1.04	1.72	0.83	0.73	1.59	5.78	1.33
RARE	**0.35**	0.94	**1.67**	**0.69**	**0.59**	**1.48**	**4.03**	**1.05**

Table 3. **Cross-dataset depth evaluation on *Car* category of KITTI and nuScenes validation (Val.) sets.** Models are trained on the KITTI training split and evaluated by depth MAE (lower is better) on the KITTI val. set (within-dataset reference) and the nuScenes frontal val. set (cross-dataset generalization). Best and second-best results are shown in bold and underlined, respectively.

Results on *Pedestrian* and *Cyclist* Categories of KITTI Test Set. In Tab. 2, we report $\mathbf{AP_{3D|40}}$, the standard metric for 'Pedestrian' and 'Cyclist' on the KITTI test set. For 'Pedestrian', RARE achieves better or comparable performance across all difficulty levels, with relative

Learn to Rank	Learn to Retrieve	Easy	Mod.	Hard
		24.93	19.04	16.57
✓		26.54	21.15	18.20
	✓	27.81	20.64	17.56
✓	✓	**28.58**	**22.05**	**19.21**

Table 4. **Overall ablations on ranking and retrieval.** *Learn to Rank* consists of point-wise and pair-wise ranking-based confidence losses, while *Learn to Retrieve* comprises query set construction, learning and retrieval.

Methods	Pearson	Spearman	Easy	Mod.	Hard
Baseline	0.540	0.480	24.93	19.04	16.57
Point-wise only	**0.827**	0.650	26.48	20.15	17.06
Pair-wise only	0.820	0.719	25.02	20.51	17.81
Ours	0.825	**0.754**	**26.54**	**21.15**	**18.20**

Table 5. **Ablations on confidence learning.** Four different methods are compared in terms of confidence estimation and 3D object detection (AP_{3D}). Confidence estimation is evaluated by the linear (Pearson) and rank (Spearman) correlations between estimated and ground truth confidence values. Higher is better.

gains of **4.4%** and **3.4%** over the second-best monocular method, MonoUNI, on the Moderate and Hard levels, respectively. The gains are most pronounced for 'Cyclist', where RARE surpasses the second-best method with large relative improvements of **52.2%/39.3%/39.7%** on the Easy/Moderate/Hard levels, respectively. Remarkably, RARE also effectively narrows the gap to LiDAR-augmented approaches such as MonoTAKD† across both categories. These results highlight that the learning to rank-and-retrieve framework is particularly well-suited to scenarios with large depth ambiguity caused by small-scale or slender objects, where single-point 3D box regression often becomes unreliable.

Cross-dataset Evaluation. Tab. 3 reports depth mean absolute error (MAE) for models trained solely on the KITTI training split and then frozen. We evaluate each model both on the KITTI validation set (within-dataset for reference) and on the nuScenes frontal validation split (cross-dataset generalization). RARE achieves the best or comparable MAE across all distance ranges on both datasets, overall outperforming recent monocular 3D detectors, indicating that it learns depth-aware representations that transfer well under dataset shift, even without any dedicated depth-specific component. Notably, RARE even surpasses MonoCD, which explicitly leverages multiple depth cues. This suggests that explicitly modeling the multimodal nature by ranking and retrieving multiple plausible 3D hypotheses can be more effective than relying solely on depth-specific components.

4.3. Ablation Studies

In this subsection, we further analyze the effectiveness of key components. Our baseline employs a single-point prediction (*i.e.*, one candidate per RoI) and uses a depth uncertainty-based confidence score, following conventional schemes such as MonoDETR [67]. Unless specified otherwise, we set the number of queries per object to three and report results for the 'Car' category on the KITTI validation set using the $AP_{3D|40}$ metric (denoted as AP_{3D} for brevity).

Effectiveness of Each Component. Tab. 4 analyzes the contribution of ranking (Sec. 3.1) and retrieval (Sec. 3.2). First, even though the model still outputs a single-point estimate per object, adding ranking-based confidence learning, jointly optimizing point-wise and pair-wise losses, already yields clear gains over the baseline, whose depth-uncertainty–based confidence is only weakly correlated with true 3D quality, as shown in Fig. 3a. This shows that reshaping scores to respect the 3D IoU-based ordering alone can substantially improve box selection under the same geometric predictions. Second, enabling retrieval without the ranking loss encourages the network to generate diverse, plausible 3D hypotheses and again surpasses the baseline, despite still relying on the weak depth uncertainty-based confidence for selection.

Finally, our full model, RARE, which combines both ranking and retrieval, achieves the best performance with relative gains of **14.6%/15.8%/15.9%** over the baseline on the Easy/Moderate/Hard levels, respectively. The improvement is larger than that of either component alone, indicating a strong synergy: the ranking loss provides reliable confidence to choose among candidates, while the retrieval module supplies rich, multimodal hypotheses for the ranking to act on. Together, they transform a single-point, uncertainty-scored predictor into a framework that can effectively sort and retrieve multiple plausible 3D boxes.

Learning to Rank. To further analyze the effect of ranking-based confidence learning (*i.e.*, Eq. 2) without retrieval, we compare four variants of confidence: depth-uncertainty (baseline), point-wise only, pair-wise only, and ours (joint point- and pair-wise learning), using three diagnostics: Pearson (linear calibration), Spearman (monotonic ranking), and AP_{3D} in Tab. 5. The baseline shows only moderate correlation with true 3D IoU and relatively low AP_{3D}. Introducing only the point-wise loss substantially improves Pearson correlation (0.540 to 0.827) and yields clear AP_{3D} gains, but its ranking alignment (Spearman) remains weaker than the pair-wise counterpart. Conversely, using only the pair-wise loss leads to stronger monotonic agreement with IoU (Spearman 0.719) and higher AP_{3D} than the baseline, at the cost of slightly reduced linear calibration. Our full variant, which jointly optimizes both, achieves the best balance, with near-optimal Pearson (0.825), the highest Spearman (0.754), and consistently superior AP_{3D} across Easy/Moderate/Hard. This confirms that combining point-wise calibration and pair-

Methods	Easy	Mod.	Hard
Baseline	24.93	19.04	16.57
Naïve learning	25.91	19.85	16.86
Retrieval learning	26.44	19.99	17.00
Retrieval learning loss w/ reg.	**27.81**	**20.64**	**17.56**

Table 6. **Ablations on retrieval learning.** Four different types of objectives are compared using 3D localization quality (AP_{3D}).

wise ordering is crucial for learning confidence that is both well-calibrated and rank-faithful.

Learning to Retrieve. To verify the effect of retrieval (*i.e.*, Eq. 8), we use the same depth uncertainty-based confidence as the baseline and vary only the training loss for candidate queries, as presented in Tab. 6. We first apply a naïve query set objective, where all queries for an object are supervised equally against the same ground truth, similar to group-based training in prior work [8]. The result is slightly better than the baseline, as different query initializations can produce mildly diverse outputs. But it provides no explicit incentive to maintain diversity and thus remains vulnerable to collapse, where all candidates converge to nearly identical boxes. We then adopt our retrieval learning, corresponding to the first term in Eq. 8, which performs retrieval-aware supervision over multiple hypotheses and further improves AP_{3D}. However, the model may still under-exploit the multimodal nature of 3D geometry since a single query can dominate the learning signal. Finally, adding the second regularization term, which explicitly encourages diverse yet plausible candidates, yields the best performance. This progression shows that simply increasing the number of queries is insufficient; retrieval-aware supervision together with diversity regularization is crucial for turning multiple candidates into complementary 3D hypotheses.

Geometric Diversity. Our learning-to-retrieve loss in Eq. 8 induces a mixture-like behavior, where multiple hypotheses compete via responsibility weights and specialize in different regions of the solution space. Fig. 4 shows the pair-wise dissimilarity among hypotheses as a function of distance. Geometric diversity increases with distance for depth, orientation, and IoU, while remaining relatively stable for height. This pattern reflects the underlying projective geometry: depth is highly sensitive to pixel-level variations at long range, whereas height is more constrained by physical priors. Moreover, the depth range covered by multiple hypotheses (e.g., $0.45 \times 3 = 1.35$ m at 40 m) closely matches the depth variation induced by pixel quantization (e.g., $\nabla D = 1.43$ m at 40 m). This suggests that the learned hypotheses adapt to the intrinsic ambiguity of the imaging process, producing diverse yet physically meaningful candidates that cover plausible 3D configurations.

Computational Efficiency. Tab. 7 compares computational efficiency of RARE against recent DETR-based methods. RARE offers a favorable accuracy–efficiency trade-off: it has the smallest model size, significantly outperforming the other methods on Moderate difficulty on the KITTI test set for *Car* category, and adds only a modest runtime overhead compared with MonoDETR. Notably, RARE is both smaller and slightly faster, yet more accurate compared with the most recent method, MonoDGP.

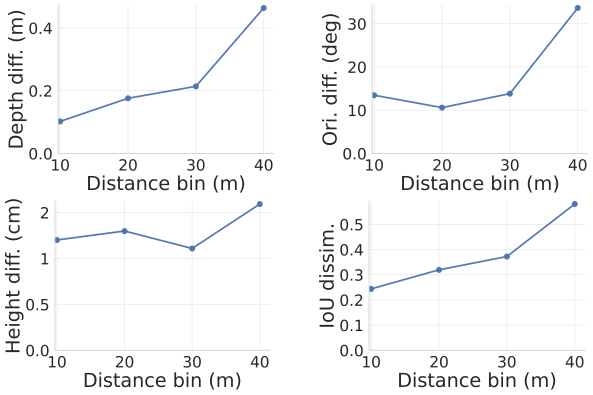


Figure 4. **Geometric diversity** of multi-hypothesis predictions across distance.

	Test, Mod. AP_{3D}	Params. (M)	Runtime (ms)
MonoDETR [67]	16.47	35.9	**28.9**
MonoDGP [43]	18.72	38.9	36.5
Ours	**19.57**	**32.9**	35.3

Table 7. **Computational efficiency and accuracy comparison.** Model size (Params.) and per-image inference time are measured with each model's official code on a single A100 GPU, averaged over the KITTI validation set, while AP_{3D} is reported on the KITTI test set for the *Car* category. Lower is better for Params. and runtime, and higher is better for AP_{3D}.

5. Conclusion

In this work, we present RARE, a novel framework that tackles the fundamental challenges of confidence estimation and 3D localization in monocular 3D object detection through a unified ranking and retrieval approach. By reformulating confidence estimation as a ranking problem rather than absolute regression, RARE produces more robust and stable confidence scores that better reflect true localization quality. Combined with its ability to predict and retrieve from multiple diverse 3D configurations, RARE explicitly addresses the inherent multimodal uncertainty in monocular 3D perception. Our extensive experimental results validate the effectiveness of this approach.

Acknowledgements. This work was supported in part by National Science Foundation (NSF) grants ECCS-2400900 and IIS-2442540, the National Artificial Intelligence Research Resource (NAIRR) Pilot, Amazon Web Services (AWS) provided through CloudBank, and NCSA Delta GPU resources provided through ACCESS.

References

[1] Christopher M Bishop. Mixture density networks. 1994. 2, 5

[2] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006. 5

[3] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9287–9296, 2019. 2

[4] Garrick Brazil, Gerard Pons-Moll, Xiaoming Liu, and Bernt Schiele. Kinematic 3d object detection in monocular video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 135–152. Springer, 2020. 2

[5] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005. 3

[6] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 5

[7] Kean Chen, Jianguo Li, Weiyao Lin, John See, Ji Wang, Lingyu Duan, Zhibo Chen, Changwei He, and Junni Zou. Towards accurate one-stage object detection with ap-loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5119–5127, 2019. 3

[8] Qiang Chen, Xiaokang Chen, Jian Wang, Shan Zhang, Kun Yao, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, and Jingdong Wang. Group detr: Fast detr training with group-wise one-to-many assignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6633–6642, 2023. 2, 8

[9] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew G Berneshawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals for accurate object class detection. *Advances in neural information processing systems*, 28, 2015. 5

[10] Wonhyeok Choi, Mingyu Shin, and Sunghoon Im. Depth-discriminative metric learning for monocular 3d object detection. *Advances in Neural Information Processing Systems*, 36, 2023. 2, 6

[11] Muhammad Sohail Danish, Muhammad Haris Khan, Muhammad Akhtar Munir, M Saquib Sarfraz, and Mohsen Ali. Improving single domain-generalized object detection: A focus on diversification and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17732–17742, 2024. 6

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 5, 6

[13] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. In *European Conference on Computer Vision*, pages 88–104. Springer, 2020. 2

[14] Yu Hong, Hang Dai, and Yong Ding. Cross-modality knowledge distillation network for monocular 3d object detection. In *European Conference on Computer Vision*, pages 87–104. Springer, 2022. 6

[15] Kuan-Chih Huang, Tsung-Han Wu, Hung-Ting Su, and Winston H Hsu. Monodtr: Monocular 3d object detection with depth-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4012–4021, 2022. 2, 6

[16] Jinrang Jia, Zhenjia Li, and Yifeng Shi. Monouni: A unified vehicle and infrastructure-side monocular 3d object detection network with sufficient depth clues. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 6

[17] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 784–799, 2018. 3

[18] Rawal Khirodkar, Visesh Chari, Amit Agrawal, and Ambrish Tyagi. Multi-instance pose networks: Rethinking top-down pose estimation. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 3122–3131, 2021. 2

[19] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[20] Abhinav Kumar, Garrick Brazil, Enrique Corona, Armin Parchami, and Xiaoming Liu. Deviant: Depth equivariant network for monocular 3d object detection. In *European Conference on Computer Vision*, pages 664–683. Springer, 2022. 1, 2, 6

[21] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9887–9895, 2019. 2

[22] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13147–13156, 2022. 2

[23] Xiang Li, Wenhai Wang, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11632–11641, 2021. 3

[24] Yingyan Li, Yuntao Chen, Jiawei He, and Zhaoxiang Zhang. Densely constrained depth estimator for monocular 3d object detection. In *European Conference on Computer Vision*, pages 718–734. Springer, 2022. 2

[25] Zhuoling Li, Zhan Qu, Yang Zhou, Jianzhuang Liu, Haoqian Wang, and Lihui Jiang. Diversity matters: Fully exploiting depth clues for reliable monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2791–2800, 2022. 2, 4, 6

[26] Zhenjia Li, Jinrang Jia, and Yifeng Shi. Monolss: Learnable sample selection for monocular 3d detection. In *2024 International Conference on 3D Vision (3DV)*, pages 1125–1135. IEEE, 2024. 2, 6

[27] Qing Lian, Peiliang Li, and Xiaozhi Chen. Monojsg: Joint semantic and geometric cost volume for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1070–1079, 2022. 6

[28] Qing Lian, Botao Ye, Ruijia Xu, Weilong Yao, and Tong Zhang. Exploring geometric consistency for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1685–1694, 2022. 6

[29] Hongbin Lin, Yifan Zhang, Shuaicheng Niu, Shuguang Cui, and Zhen Li. Monotta: Fully test-time adaptation for monocular 3d object detection. In *European Conference on Computer Vision*, pages 96–114. Springer, 2024. 2

[30] Hou-I Liu, Christine Wu, Jen-Hao Cheng, Wenhao Chai, Shian-Yun Wang, Gaowen Liu, Hugo Latapie, Jhih-Ciang Wu, Jenq-Neng Hwang, Hong-Han Shuai, et al. Monotakd: Teaching assistant knowledge distillation for monocular 3d object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22266–22275, 2025. 1, 2, 6

[31] Ji Liu, Dong Li, Rongzhang Zheng, Lu Tian, and Yi Shan. Rankdetnet: Delving into ranking constraints for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 264–273, 2021. 3, 4

[32] Xianpeng Liu, Nan Xue, and Tianfu Wu. Learning auxiliary monocular contexts helps monocular 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1810–1818, 2022. 2, 6

[33] Zongdai Liu, Dingfu Zhou, Feixiang Lu, Jin Fang, and Liangjun Zhang. Autoshape: Real-time shape-aware monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15641–15650, 2021. 2

[34] Yan Lu, Xinzhu Ma, Lei Yang, Tianzhu Zhang, Yating Liu, Qi Chu, Junjie Yan, and Wanli Ouyang. Geometry uncertainty projection network for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3111–3121, 2021. 1, 2, 3, 6

[35] Xinzhu Ma, Yinmin Zhang, Dan Xu, Dongzhan Zhou, Shuai Yi, Haojie Li, and Wanli Ouyang. Delving into localization errors for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4721–4730, 2021. 1, 2, 4, 6

[36] Xinzhu Ma, Yongtao Wang, Yinmin Zhang, Zhiyi Xia, Yuan Meng, Zhihui Wang, Haojie Li, and Wanli Ouyang. Towards fair and comprehensive comparisons for image-based 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6425–6435, 2023. 2

[37] Johannes Meier, Louis Inchingolo, Oussema Dhaouadi, Yan Xia, Jacques Kaiser, and Daniel Cremers. Monoct: Overcoming monocular 3d detection domain shift with consistent teacher models. *IEEE International Conference on Robotics and Automation (ICRA)*, 2025. 2

[38] Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan. Rank & sort loss for object detection and instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3009–3018, 2021. 3, 4

[39] Liang Peng, Fei Liu, Zhengxu Yu, Senbo Yan, Dan Deng, Zheng Yang, Haifeng Liu, and Deng Cai. Lidar point cloud guided monocular 3d object detection. In *European Conference on Computer Vision*, pages 123–139. Springer, 2022. 2, 6

[40] Liang Peng, Xiaopei Wu, Zheng Yang, Haifeng Liu, and Deng Cai. Did-m3d: Decoupling instance depth for monocular 3d object detection. In *European Conference on Computer Vision*, pages 71–88. Springer, 2022. 2, 6

[41] Liang Peng, Senbo Yan, Chenxi Huang, Xiaofei He, and Deng Cai. Digging into output representation for monocular 3d object detection, 2022. 4

[42] Liang Peng, Junkai Xu, Haoran Cheng, Zheng Yang, Xiaopei Wu, Wei Qian, Wenxiao Wang, Boxi Wu, and Deng Cai. Learning occupancy for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10281–10292, 2024. 2, 6

[43] Fanqi Pu, Yifan Wang, Jiru Deng, and Wenming Yang. Monodgp: Monocular 3d object detection with decoupled-query and geometry-error priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6520–6530, 2025. 1, 2, 6, 8

[44] Yifan Pu, Weicong Liang, Yiduo Hao, Yuhui Yuan, Yukang Yang, Chao Zhang, Han Hu, and Gao Huang. Rank-detr for high quality object detection. *Advances in Neural Information Processing Systems*, 36:16100–16113, 2023. 3

[45] Zequn Qin and Xi Li. Monoground: Detecting monocular 3d objects from the ground. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3793–3802, 2022. 1

[46] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8555–8564, 2021. 2, 6

[47] Guangsheng Shi, Ruifeng Li, and Chao Ma. Pillarnet: Real-time and high-performance pillar-based 3d object detection. In *European conference on computer vision*, pages 35–52. Springer, 2022. 3

[48] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10529–10538, 2020. 3

[49] Xuepeng Shi, Qi Ye, Xiaozhi Chen, Chuangrong Chen, Zhixiang Chen, and Tae-Kyun Kim. Geometry-based distance decomposition for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15172–15181, 2021. 4, 6

[50] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1991–1999, 2019. 2, 3, 6

[51] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Peter Kontschieder, and Elisa Ricci. Are we missing confidence in pseudo-lidar methods for monocular 3d object detection? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3225–3233, 2021. 2, 3

[52] Stephen M Stigler. Regression towards the mean, historically considered. *Statistical methods in medical research*, 6(2): 103–114, 1997. 5

[53] He Wang, Yezhen Cong, Or Litany, Yue Gao, and Leonidas J Guibas. 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14615–14624, 2021. 3

[54] Li Wang, Liang Du, Xiaoqing Ye, Yanwei Fu, Guodong Guo, Xiangyang Xue, Jianfeng Feng, and Li Zhang. Depth-conditioned dynamic message propagation for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 454–463, 2021. 2

[55] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 913–922, 2021. 1

[56] Tai Wang, Jiangmiao Pang, and Dahua Lin. Monocular 3d object detection with depth from motion. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[57] Zizhang Wu, Yunzhe Wu, Jian Pu, Xianzhi Li, and Xiaoquan Wang. Attention-based depth distillation with 3d-aware positional encoding for monocular 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2892–2900, 2023. 2

[58] Zizhang Wu, Yuanzhu Gan, Yunzhe Wu, Ruihao Wang, Xiaoquan Wang, and Jian Pu. Fd3d: Exploiting foreground depth map for feature-supervised monocular 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6189–6197, 2024. 2, 6

[59] Dongli Xu, Jinhong Deng, and Wen Li. Revisiting ap loss for dense object detection: Adaptive ranking pair selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14187–14196, 2022. 3

[60] Junkai Xu, Liang Peng, Haoran Cheng, Hao Li, Wei Qian, Ke Li, Wenxiao Wang, and Deng Cai. Mononerd: Nerf-like representations for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6814–6824, 2023. 2, 6

[61] Longfei Yan, Pei Yan, Shengzhou Xiong, Xuanyu Xiang, and Yihua Tan. Monocd: Monocular 3d object detection with complementary depths. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10248–10257, 2024. 2, 6

[62] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Std: Sparse-to-dense 3d object detector for point cloud. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1951–1960, 2019. 3

[63] Feyza Yavuz, Baris Can Cam, Adnan Harun Dogan, Kemal Oksuz, Emre Akbas, and Sinan Kalkan. Bucketed ranking-based losses for efficient training of object detectors. In *European Conference on Computer Vision*, pages 93–109. Springer, 2024. 3

[64] Qi Ye and Tae-Kyun Kim. Occlusion-aware hand pose estimation using hierarchical mixture density network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–817, 2018. 2

[65] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2018. 5

[66] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. Varifocalnet: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8514–8523, 2021. 3

[67] Renrui Zhang, Han Qiu, Tai Wang, Ziyu Guo, Ziteng Cui, Yu Qiao, Hongsheng Li, and Peng Gao. Monodetr: Depth-guided transformer for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9155–9166, 2023. 1, 2, 6, 7, 8

[68] Jinjing Zhao, Fangyun Wei, and Chang Xu. Hybrid proposal refiner: Revisiting detr series from the faster r-cnn perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17416–17426, 2024. 2

[69] Yunsong Zhou, Hongzi Zhu, Quan Liu, Shan Chang, and Minyi Guo. Monoatt: Online monocular 3d object detection with adaptive token transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17493–17503, 2023. 4

[70] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2020. 5

[71] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6748–6758, 2023. 2