Modeling and Learning Multiple Hypotheses for Monocular 3D Object Detection

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

Abstract

Detecting objects in 3D space using a monocular image is inherently a highly ill-posed problem: multiple plausible 3D bounding boxes can explain the same 2D observation of an object. Existing approaches typically follow a single-point prediction paradigm, failing to capture this multimodal nature and often regressing to an implausible mean solution. This paper introduces MonoMH, a novel multi-hypothesis framework for monocular 3D object detection. By explicitly modeling and learning the multimodal distribution of plausible 3D object configurations, MonoMH not only significantly improves detection performance but also provides richer information to support downstream decision-making. MonoMH introduces three key innovations: (1) a novel multi-hypothesis predictor that leverages spatially diverse features across different windows within an RoI to generate a rich variety of hypotheses without increasing model complexity; (2) a new multi-hypothesis learning approach that derives diverse and relevant hypotheses from single-modal ground truth by integrating uncertainty modeling with Best-of-Many learning; and (3) a hypothesis filtering mechanism that enhances detection capability by dynamically retaining a variable number of plausible hypotheses based on each object's uncertainty. Experimental results demonstrate the effectiveness of our approach. Notably, MonoMH achieves 29.12/20.88/17.93 Car AP_{3D} (easy/mod./hard) on the KITTI test set, significantly outperforming previous state-of-the-art methods. The code can be found at https://github.com/HyeonjeongPark37/MonoMH.

1. Introduction

Monocular 3D object detection aims to detect objects in 3D space using a monocular image. It holds significant promise for a wide range of applications, including autonomous driving, robotics, and extended reality. Compared with multi-view [8, 39, 53] or LiDAR-based [26, 28, 65] 3D object detection, the monocular setting offers a cost-effective and widely deployable solution but faces substantial challenges due to the lack of explicit 3D information. Modern monocular 3D object detectors [25, 32, 41, 46, 68] are built on deep neural networks that learn strong semantic and geometric representations from images, using advanced architectures tailored to this task. Recent methods [34, 51] further integrate physics-based geometric modeling with these data-driven detectors to improve performance.

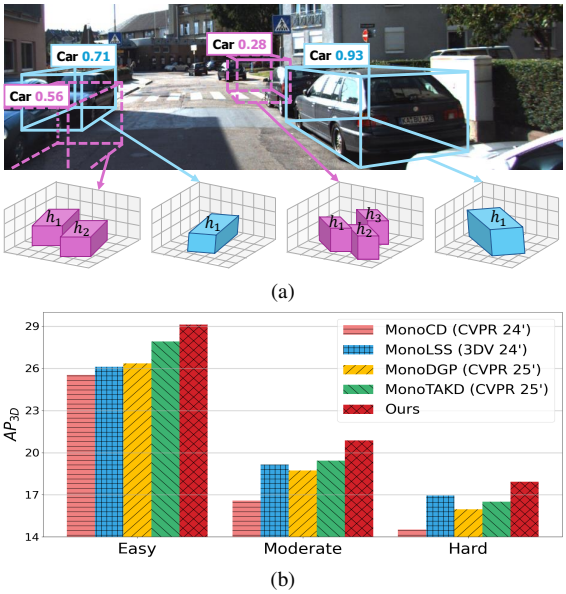

Figure 1. (a) **Visualization of multiple hypotheses predicted by MonoMH** for objects with varying levels of confidence. MonoMH effectively captures the multimodal distribution of plausible 3D bounding boxes for uncertain objects (purple), especially in scenarios of occlusion or truncation. (b) **Quantitative comparison of MonoMH against state-of-the-art methods on 'Car' category on the KITTI test set** across three difficulty levels.

Despite notable progress, monocular 3D object detection remains challenging due to its *multimodal nature*: *multiple plausible 3D bounding boxes* can explain the same 2D observation of an object, even after eliminating infeasible solutions through geometric modeling. For example, prior works [41, 51] infer depth using approximate vehicle height and the ground plane assumption. Simple derivations (in the supplementary material) reveal that even a one-pixel

change in image space can result in a wide range of depth variations in the 3D world. Nevertheless, existing methods typically model a single 3D bounding box per object. Since this *single-point prediction* inherently fails to capture the multimodal nature of monocular 3D object detection, it may regress to a mean estimate that does not correspond to any plausible solution. In practice, relying solely on such single-point predictions can pose safety risks in downstream applications like autonomous driving or robotics, where understanding the full range of 3D configurations is critical.

This paper introduces a novel multi-hypothesis framework, MonoMH, for monocular 3D object detection. As illustrated in Fig. 1, unlike traditional approaches that rely on single-point prediction, MonoMH models and learns multiple plausible 3D bounding boxes per 2D object, offering two significant advantages. First, by explicitly capturing the multimodal distribution of 3D object configurations, MonoMH overcomes the fundamental limitations of single-point prediction, which often collapses into an implausible mean estimate, and drastically enhances detection performance. Second, MonoMH provides richer information about possible 3D configurations of uncertain objects, aiding downstream applications (*e.g.*, autonomous driving and robotics) in making more informed decisions.

MonoMH addresses three core challenges in multi-hypothesis monocular 3D object detection:

(1) Multi-hypothesis Prediction. MonoMH introduces *RoI Division* to generate multiple hypotheses from each 2D detection. It divides a region of interest (RoI) into uniformly distributed windows and aggregates dense estimates from each window to propose multiple plausible 3D bounding boxes. Compared with naively using multiple detection heads, RoI Division leverages spatially diverse features across different windows to yield a richer variety of hypotheses without the burden of increased model complexity.

(2) Multi-hypothesis Learning. MonoMH learns multi-hypothesis prediction from single-modal ground truth by integrating uncertainty modeling with Best-of-Many learning. Unlike conventional uncertainty-based methods that focus on single-point learning, MonoMH propagates instance-level 3D bounding box supervision across all windows within an RoI based on their uncertainties. This ensures that the generated hypotheses are both diverse and relevant.

(3) Hypothesis Filtering. While the top hypothesis is often the closest match to the true 3D bounding box for confident objects, multiple hypotheses may be equally plausible for highly uncertain objects, such as those that are distant or heavily occluded. To address this, MonoMH introduces a hypothesis filtering mechanism that dynamically retains a varying number of hypotheses based on each object's uncertainty. This approach increases the likelihood of including correct 3D bounding boxes for uncertain objects without significantly degrading the precision.

Note that while uncertainty has been commonly used in existing monocular 3D object detectors [10, 31, 32, 41, 46] for 3D confidence calibration, all these methods only output a single 3D box prediction. DID-M3D [46] and MonoDDE [31] model pixel or keypoint-level depth estimates, but these estimates are eventually fused into a single depth prediction, and they rely on LiDAR supervision or keypoints detection. In contrast, MonoMH adaptively outputs a variable number of 3D box predictions based on the object's confidence level, without requiring additional supervision or detection. Comprehensive experimental results validate the effectiveness and applicability of MonoMH.

The contributions of this paper are summarized below:

- We present a novel multi-hypothesis framework, MonoMH, for monocular 3D object detection. Substantially different from the conventional single-point prediction paradigm, MonoMH explicitly models and learns the multimodal distribution of plausible 3D object configurations, thereby not only significantly improving detection performance but also providing richer information to support downstream decision-making.
- We propose a novel multi-hypothesis predictor that leverages spatially diverse features across different windows within an RoI to generate a rich variety of hypotheses without increasing model complexity.
- We introduce a new multi-hypothesis learning and selection approach. The learning component integrates uncertainty modeling with the Best-of-Many objective to learn diverse, relevant hypotheses from single-modal ground truth. The filtering component improves the detection capability by retaining a variable number of plausible hypotheses per object according to its predicted uncertainty.
- Extensive experimental results confirm the contributions of our approach. Notably, MonoMH consistently demonstrates strong performance across diverse baselines and benchmarks, including the KITTI test set.

2. Related Work

Monocular 3D object detection extends 2D detection to 3D, introducing new challenges due to the additional dimension. To address this, MonoCon [37] uses auxiliary tasks for 3D context, and MonoDLE [42] improves localization through better 2D-3D box alignment. GUPNet [41] introduces uncertainty estimation and hierarchical learning, while DEVIANT [25] employs scale-equivariant steerable networks. DDML [10] adds a geodesic metric loss, and MonoCD [64] decouples multi-depth prediction via complementary depths. MonoUNI [19] unifies vehicle- and infrastructure-side detection with normalized depth targets and cube-depth supervision. FD3D [62] exploits auxiliary supervision and foreground depth maps. MonoLSS [32] suggests adaptive positive sample selection with the MixUp3D augmentation. On the other hand,

MonoTTA [35] adapts at test time via reliability-driven batch normalization updates, while MonoCT [43] performs unsupervised cross-domain adaptation with depth enhancement and pseudo-label scoring. Meanwhile, detection transformers (DETR) [5, 72] have been extended to 3D in several studies [16, 38, 49, 68, 71]. MonoDETR [68] proposes a depth-aware transformer, while MonoXiver [38] bridges 2D-to-3D and 3D-to-2D mappings using the Perceiver I/O model [17]. MonoATT [71] improves online efficiency via adaptive tokenization, and MonoDGP [49] models geometry error with a decoupled decoder.

Some approaches alleviate the lack of depth input by incorporating additional training data, such as LiDAR point clouds [16, 36, 45, 48, 52, 63], CAD models [30, 40], videos [4, 60], and depth data [46, 58, 61]. These works typically follow the single-point prediction paradigm. A few methods [3, 47] predict multiple 3D boxes per 2D object, but they rely on predefined fixed anchors and do not perform any multi-hypothesis learning. In contrast, MonoMH learns to dynamically generate variable distributions of 3D boxes based on the uncertainty of each observed object.

Multi-hypothesis prediction has been explored in areas such as 3D human pose estimation [18, 23, 27] and hand pose estimation [6, 66], primarily to address the ambiguity in recovering 3D information from monocular input. For example, some methods [27, 66], based on Mixture Density Networks (MDN) [2], sample multiple 3D pose hypotheses consistent with a given 2D pose by utilizing a mixture of Gaussians. On the other hand, adding multiple detection heads [23] enables a one-to-many mapping, and a recent architecture [29] adds many-to-one relations across hypotheses to refine the final 3D pose. These multi-hypothesis techniques have proven highly effective in overcoming the multimodal depth ambiguity inherent in inverse problems. Nevertheless, multi-hypothesis modeling and learning in monocular 3D object detection remain unexplored.

3. Method

Monocular 3D object detection estimates 3D boxes from a single RGB image. This task remains challenging due to its multimodal nature, where multiple plausible 3D bounding boxes may explain the same 2D observation of an object. We introduce MonoMH, a novel multi-hypothesis framework for monocular 3D object detection. Unlike traditional single-point prediction, which often collapses into an implausible mean estimate, MonoMH explicitly models and learns the multimodal distribution of plausible 3D object configurations, *i.e.*, multiple 3D bounding box hypotheses and their uncertainties. Consequently, MonoMH not only improves detection performance but also provides richer information to support downstream decision-making.

MonoMH builds upon a base detector and introduces three key innovations: multi-hypothesis prediction, multi-hypothesis learning, and hypothesis filtering. We first briefly describe the base detector below and then elaborate on our main contributions in Sec. 3.1-Sec. 3.3.

Base Detector. We adopt a common architecture used in recent state-of-the-art methods [9, 32, 41, 46, 64, 69, 70] as our generic base detector. A backbone network (*i.e.*, DLA34 [67]) extracts image features, and 2D detection heads predict an object class distribution, a 2D box size, and a 2D offset at each position of the feature map. Finally, 3D heads lift each 2D detection to a 3D box by regressing the object's depth, uncertainty, orientation, center projection offset, and 3D size. Following prior work [32, 41, 46], we model the Laplacian aleatoric uncertainty, which is parametrized as the standard deviation of a Laplace distribution. The 2D and 3D heads are trained by comparing their predictions with the corresponding ground truth, weighted by uncertainty estimates [32, 41, 46]. Further details are provided in the supplementary material.

3.1. Multi-hypothesis Prediction

A straightforward approach for multi-hypothesis prediction is to apply multiple 2D-to-3D lifting heads to the RoI features, with each head predicting a separate hypothesis. However, this method has two critical limitations. First, it lacks scalability: both model size and computational complexity increase with the number of hypotheses. Second, as each head shares the same input, *i.e.*, the RoI features, the learning process must be carefully designed to ensure diversity among hypotheses. To overcome these limitations, we propose a fundamentally different solution: RoI Division, which is added to the base detector as a separate branch.

Let $\boldsymbol{F}^{\text{roi}} \in \mathbb{R}^{M \times M \times C}$ denote the RoI feature map pooled from a 2D object detection, where $M \times M$ is the pooled region size (typically 7×7) and C is the feature dimension. This RoI feature map is fed into parallel branches, each comprising two convolutional layers, to predict a depth map $\boldsymbol{D}^{\text{roi}} \in \mathbb{R}^{M \times M}$, an uncertainty map $\boldsymbol{U}^{\text{roi}} \in \mathbb{R}^{M \times M}$, a 3D size $\boldsymbol{s}^{\text{roi}} \in \mathbb{R}^3$, a 3D center projection offset $\boldsymbol{o}^{\text{roi}} \in \mathbb{R}^2$, and an orientation θ^{roi}. We then convert the uncertainty map $\boldsymbol{U}^{\text{roi}}$ to a confidence map $\boldsymbol{C}^{\text{roi}} \in [0,1]^{M \times M}$ using a sigmoid function: $\boldsymbol{C}^{\text{roi}} = \sigma(-\boldsymbol{U}^{\text{roi}})$. This transformation ensures that high uncertainty corresponds to low confidence, and low uncertainty corresponds to high confidence.

RoI Division generates K hypotheses by partitioning the 2D RoI feature map into K fixed grid of spatial windows, as illustrated in Fig. 2, and aggregating dense estimates within each window. The k-th hypothesis is represented as

$$\mathcal{H}_k = \{x_k, y_k, d_k, \boldsymbol{s}_k, \theta_k, c_k\}, k = 1, \ldots, K \qquad (1)$$

where (x_k, y_k), d_k, $\boldsymbol{s}_k$, θ_k, and c_k are the k-th hypothesis's x and y-coordinates in 3D space, depth, 3D size, orientation, and confidence, respectively.

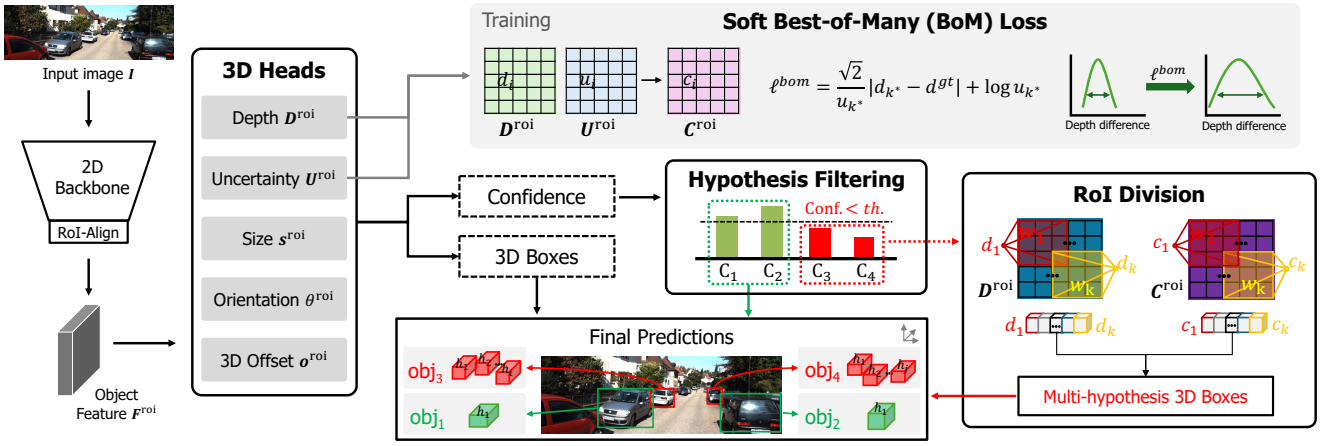

Figure 2. **MonoMH overall framework.** Different from existing methods that predict a single 3D bounding box for each 2D object detection, MonoMH learns to predict a variable number of 3D bounding boxes based on the object's confidence level. To this end, MonoMH comprises multi-hypothesis prediction via RoI Division, hypothesis filtering, and multi-hypothesis learning with a soft Best-of-Many loss. All panels depict inference except the gray 'Training' panel.

We construct $\mathcal{H}_k$ as follows:

$$x_k = \left(\frac{x^{2D} + o_x^{roi} - c_x}{f_x} \right) d_k \tag{2}$$

$$y_k = \left(\frac{y^{2D} + o_y^{roi} - c_y}{f_y} \right) d_k \tag{3}$$

$$d_k = \frac{\sum_{i \in \mathcal{W}_k} c_i^{roi} \cdot d_i^{roi}}{\sum_{i \in \mathcal{W}_k} c_i^{roi}} \tag{4}$$

$$\boldsymbol{s}_k = \boldsymbol{s}^{roi} \tag{5}$$

$$\theta_k = \theta^{roi} \tag{6}$$

$$c_k = \frac{\sum_{i \in \mathcal{W}_k} c_i^{roi} \cdot c_i^{roi}}{\sum_{i \in \mathcal{W}_k} c_i^{roi}} \tag{7}$$

Eqs. (2) and (3) calculate (x_k, y_k) based on the depth d_k, the 3D center projection offset $\boldsymbol{o}^{roi} = (o_x^{roi}, o_y^{roi})$, and the camera intrinsic matrix $\begin{bmatrix} f_x & 0 & c_x; 0 & f_y & c_y; 0 & 0 & 0 \end{bmatrix}$. Eqs. (4) and (7) calculate d_k and c_k as the confidence-weighted depth and confidence averages within $\mathcal{W}_k$, where $\mathcal{W}_k$ is the k-th window in the RoI. Here, d_i^{roi} and c_i^{roi} denote the i-th elements of $\boldsymbol{D}^{roi}$ and $\boldsymbol{C}^{roi}$, respectively. Eqs. (5) and (6) assign the RoI-based size and orientation predictions to $\boldsymbol{s}_k$ and θ_k, respectively. Since it is always the inaccurate 3D localization rather than the bounding box size or orientation that leads to incorrect detections, we allow independent localization and confidence across hypotheses while sharing bounding box size and orientation.

A remaining question is how to divide the RoI into windows. We follow three principles. First, each window should cover a part-level region, *e.g.*, 3×3 windows in a 7×7 RoI. Second, the number of windows determines the number of hypotheses: too few may miss plausible configurations, while too many can introduce redundancy. We find that 9 windows arranged in a 4×4 grid often work well. Third, the windows should have minimal overlap to encourage diversity. Our experiments in the supplementary material show that MonoMH is robust to these design choices.

Discussion. RoI Division leverages spatially diverse features from different windows to generate a rich variety of high-quality hypotheses without the burden of increased model complexity. Concretely, features from distinct parts of an object, such as the front wheels or trunk of a car, are crucial for accurate 3D localization. Each window emphasizes a particular part while still capturing global object context, thanks to the large receptive fields in deeper backbone layers. This allows RoI Division to produce relevant hypotheses from different perspectives. Even when certain parts are occluded, the visible regions can still yield reliable predictions. Moreover, since the prediction head is shared across all windows, RoI Division avoids a larger model size or a higher risk of overfitting.

3.2. Multi-hypothesis Learning

Since the box size and orientation are shared across hypotheses, they can be learned similarly to the base detector by comparing $\boldsymbol{s}^{roi}$ and θ^{roi} with their respective ground truth. The x- and y-coordinates $\{(x_k, y_k) : k = 1, \ldots, K\}$ vary across hypotheses due solely to differences in depth. Thus, the main challenge is to effectively learn the multi-hypothesis predictions for depth and their associated confidences $\{(d_k, c_k) : k = 1, \ldots, K\}$.

We draw inspiration from the Best-of-Many (BoM) loss [1], which has shown effectiveness in learning multimodal trajectory predictions from single-modal supervision. BoM selects the prediction closest to the ground truth from multiple hypotheses and minimizes the difference between this best prediction and the ground truth. To learn multi-

hypothesis predictions for both depth and confidence with only depth ground truth, it is intuitive to integrate BoM [1] and uncertainty modeling [41]:

$$\ell^{\text{bom}} = \frac{\sqrt{2}}{u_{k^*}} |d_{k^*} - d^{\text{gt}}| + \log u_{k^*} \tag{8}$$

$$k^* = \arg\min_k |d_k - d^{\text{gt}}| \tag{9}$$

where d^{gt} is the ground truth depth, k^* is the index of the best hypothesis, and u_k is the confidence-weighted uncertainty averaged within a window.

The loss function in Eq. (8) has proven effective in our experiments, but it has one potential drawback: since only the closest hypothesis is supervised, only pixels within its corresponding window receive gradients, which may limit the learning effectiveness. This motivates us to explore a variant of our loss function by relaxing the hard BoM to a *soft* BoM. Instead of selecting only the best hypothesis, we approximate d_{k^*} and u_{k^*} as confidence-weighted depth and uncertainty averages over the RoI, respectively: $d_{k^*} = \sum_i c_i^{\text{roi}} \cdot d_i^{\text{roi}} / \sum_i c_i^{\text{roi}}$ and $u_{k^*} = \sum_i c_i^{\text{roi}} \cdot u_i^{\text{roi}} / \sum_i c_i^{\text{roi}}$. Note that uncertainty is optimized only through this objective and learned end-to-end via backpropagation without extra supervision, similar to [22]. This relaxation (i) allows all pixels to receive gradients proportional to their confidence and (ii) avoids back-propagating conflicting gradients from multiple hypotheses to overlapping pixels. Empirically, the soft BoM consistently outperforms the hard BoM variant.

Overall Training Objective. Having defined depth and confidence learning (soft BoM) and the standard 2D/3D heads, we summarize the total loss. The 2D loss $\mathcal{L}^{\text{2D}}$ follows CenterNet [70] (heatmap, 2D size, 2D offset). The 3D loss $\mathcal{L}^{\text{3D}}$ uses a MultiBin orientation loss [9], Smooth-L_1 for the 3D center projection offset, L_1 for 3D size, and a pixel-wise Laplacian depth loss [32, 46]. We simply sum all terms without specific weights:

$$\mathcal{L}^{\text{total}} = \mathcal{L}^{\text{2D}} + \mathcal{L}^{\text{3D}} + \ell^{\text{bom}} \tag{10}$$

Discussion. We do not add explicit regularizers to encourage diversity, as RoI Division (Sec. 3.1) inherently induces it, as supported by our statistics (Fig. 4). In monocular 3D detection, similar 2D observations can map to very different 3D ground truths. Therefore, the BoM loss requires at least one hypothesis to align well with each single-modal ground truth, so it is minimized only when the diverse hypotheses effectively capture the multimodal distribution of 3D configurations conditioned on each 2D object. Associating each hypothesis with an uncertainty allows the detector to express confidence in different hypotheses, which is necessary for hypothesis filtering (Sec. 3.3). The soft BoM loss stabilizes training by distributing gradients across windows proportionally to their confidence, reducing conflicts.

3.3. Hypothesis Filtering

During inference, a simple way to use multi-hypothesis predictions from RoI Division is to select the highest-confidence hypothesis or use a soft best estimate (*i.e.*, confidence-weighted average). This works well for confident objects. However, for highly uncertain cases, such as distant or heavily occluded objects, confidence levels across all hypotheses tend to be low. In such scenarios, even the best hypothesis may fail to capture the true 3D configuration, and each hypothesis remains a plausible candidate.

The analysis motivates us to explore a hypothesis filtering mechanism that dynamically retains a variable number of hypotheses based on uncertainty. If the confidence of the top hypothesis is higher than a threshold, only that hypothesis is retained (*e.g.*, the green 3D boxes in Fig. 2). Otherwise, all hypotheses whose confidence values are within a specified range of the highest confidence are kept to capture multiple modes of the uncertain 3D object configuration (*e.g.*, the red 3D boxes in Fig. 2).

Discussion. Despite its simplicity, our hypothesis filtering proves both practically valuable and theoretically sound. In practice, it provides richer information about all plausible 3D configurations of uncertain objects, enabling downstream applications to make more informed decisions and improve safety. For example, recent advancements in autonomous driving [15, 50, 55, 57] and robotics [11, 13, 21, 44] have integrated uncertainty in 3D perception into planning, grasping, and manipulation algorithms for improved robustness. Characterizing the multimodal distribution of 3D detections has the potential to significantly impact these areas. Theoretically, for highly uncertain detections, retaining multiple predictions or modes with similar precision levels increases the likelihood of finding all objects of interest in 3D space without severely compromising precision. A detailed mathematical analysis and illustration are provided in the supplementary material. This enhanced detection capability further justifies the value of hypothesis filtering in practical applications.

4. Experiments

4.1. Setup

Datasets and Evaluation Metrics. We evaluate our method on KITTI [12] and Waymo [56] benchmarks.

- ***KITTI*** includes 7,481 training and 7,518 test images. Following the common split [7], we use 3,712 images for training and 3,769 for validation. The dataset defines 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' [54].
- ***Waymo*** contains 52,386 training and 39,848 validation

Methods	$AP_{3D\|40}$			$AP_{BEV\|40}$		
	Easy	Mod.	Hard	Easy	Mod.	Hard
CaDDN† [52]('21)	19.17	13.41	11.46	27.94	18.91	17.19
MonoDTR† [16]('22)	21.99	15.39	12.73	28.59	20.38	17.14
DID-M3D† [46]('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† [45]('22)	25.56	17.80	15.38	35.96	24.81	21.86
MonoNeRD† [63]('23)	22.75	17.13	15.63	31.13	23.46	20.97
OM3D† [48]('24)	25.55	17.02	14.79	35.38	24.18	21.37
MonoTAKD† [36] ('25)	27.91	19.43	16.51	38.75	27.76	24.14
MonoDLE [42]('21)	17.23	12.26	10.29	24.79	18.89	16.00
GUPNet [41]('21)	20.11	14.20	11.77	-	-	-
DEVIANT [25]('22)	21.88	14.46	11.89	29.65	20.44	17.43
MonoCon [37]('22)	22.50	16.46	13.95	31.12	22.10	19.00
MonoJSG [33]('22)	24.69	16.14	13.64	32.59	21.26	18.18
MonoDDE [31]('22)	24.93	17.14	15.10	33.58	23.46	20.37
MonoDETR [68]('23)	25.00	16.47	13.58	33.60	22.11	18.60
MonoXiver [38]('23)	25.24	19.04	16.39	34.14	25.37	22.20
DDML [10]('23)	23.31	16.36	13.73	-	-	-
MonoUNI [19]('23)	24.75	16.73	13.49	-	-	-
MonoATT [71]('23)	24.72	17.37	15.00	36.87	24.42	21.88
FD3D [62]('24)	25.38	17.12	14.50	34.20	23.72	20.76
MonoCD [64]('24)	25.53	16.59	14.53	33.41	22.81	19.57
MonoMAE [20]('24)	25.60	18.84	16.78	34.15	24.93	21.76
MonoLSS [32]('24)	26.11	19.15	16.94	34.89	25.95	22.59
MonoDGP [49] ('25)	26.35	18.72	15.97	35.24	25.23	22.02
MonoMH	**29.12**	**20.88**	**17.93**	**37.85**	**28.06**	**24.53**

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 auxiliary training data. Best and second-best results are shown in bold and underlined, respectively.

images from the front camera. Following [25, 52, 59], we sample every third frame for training and evaluate on the validation set, focusing on the 'Vehicle' class. Evaluation is conducted at Level 1 and Level 2, across three distance ranges: [0, 30), [30, 50), and [50, ∞) meters, where each level reflects LiDAR point density. We report AP_{3D} and APH_{3D}, where APH_{3D} incorporates heading accuracy.

Implementation Details. MonoMH is trained on a single Tesla V100 GPU with batch sizes of 16 for KITTI and 40 for Waymo. To stabilize training, we employ Hierarchical Task Learning (HTL) [41] with a linear warm-up strategy. For KITTI, we apply MixUp3D [32] data augmentation to address its limited training size. We use Adam optimizer [24] with an initial learning rate of 0.00125. The number of epochs is set to 600 for KITTI and 30 for Waymo. The RoI size is set to 7×7. We generate 9 hypotheses using a 4×4 window for KITTI and 5 hypotheses using a 5×5 window for Waymo. Additionally, the confidence threshold is 0.75 and 0.5 for KITTI and Waymo, respectively. We only consider hypotheses with a depth difference within two meters of the top hypothesis, which helps exclude outliers, such as those from occluded regions.

4.2. Main Results

Results on *Car* Category of KITTI Test Set. Tab. 1 compares MonoMH with recent state-of-the-art methods on the KITTI test set for the *Car* category. Under the purely monocular setting, MonoMH achieves the highest $\mathbf{AP_{3D|40}}$ and $\mathbf{AP_{BEV|40}}$ across all difficulty levels, with relative gains of **+10.51%/+9.03%/+5.84%** in $\mathbf{AP_{3D|40}}$ and **+2.66%/+8.13%/+8.58%** in $\mathbf{AP_{BEV|40}}$ over the second-best results on the Easy/Moderate/Hard levels, respectively. Moreover, despite using only monocular supervision, MonoMH also surpasses MonoTAKD, a BEV-oriented method with LiDAR-guided distillation, in $\mathbf{AP_{3D|40}}$ for all difficulty levels and in $\mathbf{AP_{BEV|40}}$ on the Moderate and Hard levels while remaining competitive on the Easy level. These results highlight the strength of our multi-hypothesis modeling in purely monocular settings.

[1]MonoLSS repository issue: link

Methods	Ped., $AP_{3D\|40}$			Cyc., $AP_{3D\|40}$		
	Easy	Mod.	Hard	Easy	Mod.	Hard
CaDDN† [52]('21)	12.87	8.14	6.76	7.00	3.41	3.30
MonoDTR† [16]('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† [48]('24)	14.68	9.15	7.80	7.37	3.56	2.84
MonoDLE [42]('21)	9.64	6.55	5.44	4.59	2.66	2.45
GUPNet [41]('21)	14.72	9.53	7.87	4.18	2.65	2.09
DEVIANT [25]('22)	13.43	8.65	7.69	5.05	3.13	2.59
MonoCon [37]('22)	13.10	8.41	6.94	2.80	1.92	1.55
MonoJSG [33]('22)	11.02	7.49	6.41	5.45	3.21	2.57
MonoDDE [31]('22)	11.13	7.32	6.67	5.94	3.78	3.33
DDML [10]('23)	14.90	10.28	8.70	5.38	2.89	2.83
MonoUNI [19]('23)	15.78	10.34	8.74	7.34	4.28	3.78
MonoLSS [32]('24)	17.09	11.27	10.00	7.23	4.34	3.92
MonoMH	**18.42**	**12.15**	**10.36**	**11.39**	**6.70**	**5.83**

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.

IoU /D.	Method	AP_{3D}				APH_{3D}			
		All	0-30	30-50	50+	All	0-30	30-50	50+
0.7 /L1	OM3D† [48]	10.61	29.18	4.49	0.41	10.53	28.96	4.46	0.40
	MonoLSS† [32]	3.71	9.82	1.14	0.16	3.69	9.75	1.13	0.16
	GUPNet [41]	2.28	6.15	0.81	0.03	2.27	6.11	0.80	0.03
	DEVIANT [25]	2.69	6.95	0.99	0.02	2.67	**6.90**	0.98	0.02
	MonoMH	**2.78**	**6.96**	**1.35**	**0.11**	**2.76**	**6.90**	**1.33**	**0.11**
0.7 /L2	OM3D† [48]	10.02	28.38	4.38	0.36	9.94	28.17	4.34	0.36
	MonoLSS† [32]	3.27	9.79	1.11	0.15	3.25	9.73	1.10	0.15
	GUPNet [41]	2.14	6.13	0.78	0.02	2.12	6.08	0.77	0.02
	DEVIANT [25]	2.52	**6.93**	0.95	0.02	2.50	6.87	0.94	0.02
	MonoMH	**2.61**	**6.93**	**1.30**	**0.10**	**2.59**	**6.88**	**1.28**	**0.10**
0.5 /L1	OM3D† [48]	28.99	61.24	23.25	3.65	28.66	60.63	23.00	3.59
	MonoLSS† [32]	13.49	33.64	6.45	1.29	13.38	33.39	6.40	1.26
	GUPNet [41]	10.02	24.78	4.84	0.22	9.94	24.59	4.78	0.22
	DEVIANT [25]	10.98	26.85	5.13	0.18	10.89	26.64	5.08	0.18
	MonoMH	**12.28**	**28.56**	**7.18**	**0.65**	**12.17**	**28.32**	**7.09**	**0.65**
0.5 /L2	OM3D† [48]	27.21	61.09	22.59	3.18	26.90	60.49	22.34	3.13
	MonoLSS† [32]	13.12	33.56	6.28	1.15	13.02	33.32	6.22	1.13
	GUPNet [41]	9.39	24.69	4.67	0.19	9.41	24.50	4.62	0.19
	DEVIANT [25]	10.29	26.75	4.95	0.16	10.20	26.54	4.90	0.16
	MonoMH	**11.51**	**28.45**	**6.92**	**0.57**	**11.40**	**28.21**	**6.83**	**0.56**

Table 3. **Comparison on the *Vehicle* category on the Waymo validation set.** 'D.' denotes difficulties (L1=Level_1, L2=Level_2). Methods marked with † use extra training data. OM3D† uses LiDAR for training. MonoLSS† has a bug in its code[1] that uses validation images for training data augmentation.

Results on *Pedestrian* and *Cyclist* Categories of KITTI Test Set. In Tab. 2, we evaluate the $AP_{3D|40}$, which is the

Architecture	Method	$AP_{3D\|40}$		
		Easy	Mod.	Hard
Conv.-based	MonoDLE [42]	17.45	13.66	11.68
	+ **Ours**	**21.77**	**16.57**	**14.34**
	Improvement	+4.32	+2.91	+2.66
	GUPNet [41]	22.76	16.46	13.72
	+ **Ours**	**24.53**	**18.63**	**16.30**
	Improvement	+1.77	+2.17	+2.58
	DEVIANT [25]	24.63	16.54	14.52
	+ **Ours**	**25.42**	**19.09**	**16.49**
	Improvement	+0.79	+2.55	+1.97
	MonoLSS [32]	25.91	18.29	15.94
	+ **Ours**	**28.07**	**20.01**	**16.80**
	Improvement	+2.16	+1.72	+0.86
DETR-based	MonoDETR [68]	28.84	20.61	16.38
	+ **Ours**	**30.25**	**22.00**	**18.54**
	Improvement	+1.41	+1.39	+2.16

Table 4. **Applicability to other methods.** To show generalizability, we integrate our approach into four Conv.-based baselines and one DETR-based method. Best results are in bold.

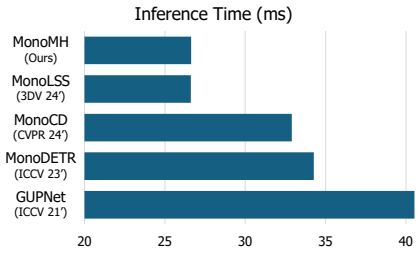

Figure 3. **Inference time comparison.** Per-image inference time is measured with each model's official code on a single V100 GPU (batch size=1), averaged over the KITTI validation set.

commonly reported evaluation metric for 'Pedestrian' and 'Cyclist' on the KITTI test set. Our method consistently outperforms all comparable approaches across all classes and difficulty levels. Especially, our method achieves significant improvements in the 'Cyclist' category. Specifically, it achieves relative improvements of **+55.2%** for Easy, **+54.4%** for Moderate, and **+48.7%** for Hard compared to the previous best-performing method.

Results on *Vehicle* Category of Waymo Validation Set. We assess generalization on the Waymo dataset, which is larger and more diverse than KITTI. Since recent papers have not released code for Waymo, we use GUPNet [41] as our strong, reproducible baseline and implement our multi-hypothesis approach on it. Tab. 3 shows that MonoMH consistently outperforms the other methods across all ranges under the monocular setting without extra training data.

Applicability to Other Methods. Tab. 4 demonstrates the effectiveness and generalizability of our approach by integrating multi-hypothesis modeling and learning into four convolution (Conv.)-based baselines and one DETR-based detector. Our method consistently yields notable performance improvements across all settings, underscoring its broad applicability. In particular, when applied to the DETR-based MonoDETR, our method adopts

Method	Easy	Mod.	Hard
Baseline	23.81	17.00	14.17
Multi-head	24.13	16.59	13.65
RoI Division (ours)	**28.94**	**21.53**	**18.43**

Table 5. **Comparison of multi-hypothesis prediction methods.**

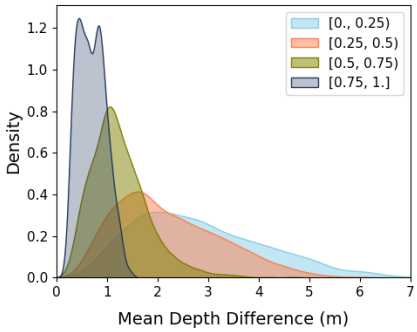

Figure 4. **Distribution of mean depth differences between hypotheses generated by RoI Division within an RoI** across four confidence ranges. RoI Division yields broader depth hypotheses at lower confidence.

content-dependent object queries inspired by Deformable DETR [72]. These improvements confirm that MonoMH is architecture-agnostic and broadly effective across diverse monocular 3D detection architectures.

Efficiency. Fig. 3 compares inference time of MonoMH and representative state-of-the-art methods. All networks are evaluated under the same computing environment. MonoMH achieves an inference speed that is faster or comparable to other methods.

4.3. Ablation Studies

In this subsection, we conduct ablations to analyze the impact of key components. The base detector in Sec. 3 serves as the baseline. Unless specified otherwise, results are reported for the 'Car' category on the KITTI validation set using the $AP_{3D|40}$ metric.

Multi-hypothesis Prediction. To assess the effectiveness of different multi-hypothesis generation strategies, we compare the conventional multi-head approach with our RoI Division. Specifically, we add nine 3D heads in the multi-head approach, whereas RoI Division generates nine hypotheses from a single 3D head. As shown in Tab. 5, our method consistently outperforms the multi-head approach across all difficulty levels, highlighting the benefit of leveraging spatially diverse features within a RoI. Fig. 4 verifies the diversity of hypotheses produced by RoI Division, despite the absence of any explicit diversity regularization during training. Depth diversity increases as confidence decreases, indicating that RoI Division yields broader depth hypotheses to reflect depth ambiguity.

Fig. 5 provides insight into the spatial behavior of hypotheses by showing the distribution of the best hypothesis,

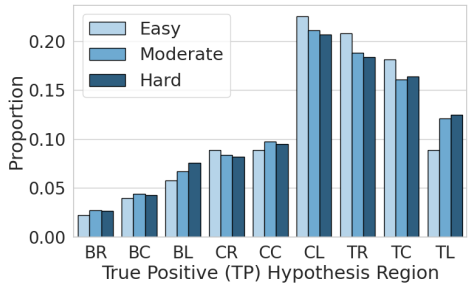

Figure 5. **Spatial distribution of the best hypothesis (*i.e.*, true positive)** across the nine window regions under the Easy/Moderate/Hard levels. Each bar indicates the proportion of true positives located in a specific region. (TL = top-left, TC = top-center, TR = top-right, CL = center-left, CC = center-center, CR = center-right, BL = bottom-left, BC = bottom-center, BR = bottom-right)

i.e., true positive (TP), across nine window regions. Most TPs are concentrated near the center, confirming that central views are typically informative. As difficulty increases, however, non-central regions gain importance. Specifically, the top-left region becomes more prominent from Easy to Moderate, and the bottom-left from Moderate to Hard. This trend underscores the importance of spatial diversity: plausible hypotheses often emerge from peripheral regions under occlusion or truncation. These findings validate our design choice to preserve multiple localized perspectives rather than rely solely on central predictions.

Multi-hypothesis Learning. Tab. 6 shows that integrating Best-of-Many loss with uncertainty modeling achieves effective multi-hypothesis learning, with the soft variant yielding additional performance gains. Additionally, we observe that supervising the depth estimate at each pixel using only the depth ground truth results in inferior performance.

Hypothesis Filtering. To validate our hypothesis filtering strategy, we compare three ways to process hypotheses from RoI Division: confidence-weighted averaging ('Mean'), selecting the highest-confidence hypothesis ('Best'), and our uncertainty-based filtering ('Filtering'), which extracts multiple hypotheses when confidence is low. As depicted in Tab. 7, all strategies surpass the baseline, confirming the general benefit of multi-hypothesis modeling and learning. Our filtering strategy yields the best results across all difficulties, indicating that keeping a small set of plausible hypotheses under uncertainty is most effective. Furthermore, Fig. 6 reports the precision–recall curve comparing MonoMH, MonoMH-best, MonoLSS, and Base (our base detector) for the 'Car' category at easy difficulty. MonoMH-best, keeping only the top hypothesis, matches or exceeds the precision of single-point prediction (MonoLSS, Base). MonoMH maintains similar precision yet lifts recall in the high-recall regime, where low-confidence detections become important, by retaining only a few hypotheses whose confidences are near the top; these capture distinct, plausible modes of uncertain 3D configurations. This strategy increases the likelihood of finding all 3D objects without significantly sacrificing precision (formal analysis is available in the supplementary material).

Method	Easy	Mod.	Hard
Baseline	23.81	17.00	14.17
Pixel-wise depth supervision	24.42	17.56	14.71
Hard Best-of-Many	26.29	19.96	17.05
Soft Best-of-Many (ours)	**28.94**	**21.53**	**18.43**

Table 6. **Comparison of loss functions for multi-hypothesis learning.**

Method	Easy	Mod.	Hard
Baseline	23.81	17.00	14.17
Mean	25.96	18.45	16.13
Best	26.89	19.62	17.27
Filtering (ours)	**28.94**	**21.53**	**18.43**

Table 7. **Comparison of hypotheses processing methods.**

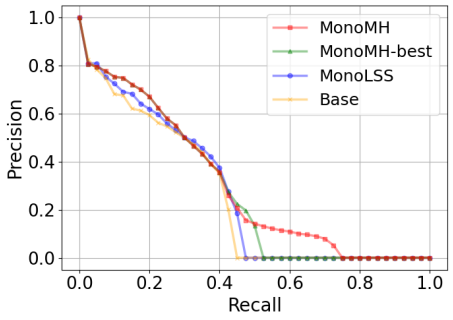

Figure 6. **Precision-recall curve** for the 'Car' category at the 'Easy' difficulty, comparing MonoMH, MonoMH-best, MonoLSS, and Base.

More results, including hyper-parameter studies, ablations of key components on other categories, and qualitative examples, are available in the supplementary material.

5. Conclusion

We introduce MonoMH, a novel multi-hypothesis modeling and learning framework for monocular 3D object detection. MonoMH combines (i) RoI Division for diverse hypothesis generation, (ii) soft Best-of-Many loss to learn multi-modal 3D object configurations, and (iii) hypothesis filtering to selectively retain hypotheses for confident and uncertain objects. Extensive experiments validate its robustness and effectiveness. As future work, we will explore better uncertainty models and adaptive hypothesis filtering to further enhance MonoMH.

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] Apratim Bhattacharyya, Bernt Schiele, and Mario Fritz. Accurate and diverse sampling of sequences based on a "best of many" sample objective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8485–8493, 2018. 4, 5

[2] Christopher M Bishop. Mixture density networks. 1994. 3

[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. 3

[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. 3

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[6] Rongyu Chen, Linlin Yang, and Angela Yao. Mhentropy: Entropy meets multiple hypotheses for pose and shape recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14840–14849, 2023. 3

[7] 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

[8] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017. 1

[9] Yongjian Chen, Lei Tai, Kai Sun, and Mingyang Li. Monopair: Monocular 3d object detection using pairwise spatial relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12093–12102, 2020. 3, 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] Viral Rasik Galaiya, Thiago Eustaquio Alves De Oliveira, Xianta Jiang, and Vinicius Prado Da Fonseca. Grasp approach under positional uncertainty using compliant tactile sensing modules and reinforcement learning. In *2024 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 424–428. IEEE, 2024. 5

[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] Suqin He, Weiye Zhao, Chuxiong Hu, Yu Zhu, and Changliu Liu. A hierarchical long short term safety framework for efficient robot manipulation under uncertainty. *Robotics and Computer-Integrated Manufacturing*, 82:102522, 2023. 5

[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] Wen Hu, Cong Wang, Zejian Deng, Yanding Yang, Yang Wu, Kai Cao, Bangji Zhang, and Dongpu Cao. Uncertainty-aware decision making and planning for icv based on asymmetric driving aggressiveness. *IEEE Transactions on Intelligent Vehicles*, 2024. 5

[16] 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. 3, 6

[17] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, 2021. 3

[18] Ehsan Jahangiri and Alan L Yuille. Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 805–814, 2017. 3

[19] 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. 2, 6

[20] Xueying Jiang, Sheng Jin, Xiaoqin Zhang, Ling Shao, and Shijian Lu. Monomae: Enhancing monocular 3d detection through depth-aware masked autoencoders. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 6

[21] Chunting Jiao, Lishuai Yu, Xiaojie Su, Yao Wen, and Xin Dai. Adaptive hybrid impedance control for dual-arm cooperative manipulation with object uncertainties. *Automatica*, 140:110232, 2022. 5

[22] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018. 5

[23] 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. 3

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

[25] 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, 7

[26] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders

for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 1

[27] 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. 3

[28] Jinyu Li, Chenxu Luo, and Xiaodong Yang. Pillarnext: Rethinking network designs for 3d object detection in lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17567–17576, 2023. 1

[29] 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. 3

[30] 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. 3

[31] 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, 6

[32] 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. 1, 2, 3, 5, 6, 7

[33] 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

[34] 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. 1

[35] 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. 3

[36] 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. 3, 6

[37] 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

[38] Xianpeng Liu, Ce Zheng, Kelvin B Cheng, Nan Xue, Guo-Jun Qi, and Tianfu Wu. Monocular 3d object detection with bounding box denoising in 3d by perceiver. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6436–6446, 2023. 3, 6

[39] Xianpeng Liu, Ce Zheng, Ming Qian, Nan Xue, Chen Chen, Zhebin Zhang, Chen Li, and Tianfu Wu. Multi-view attentive contextualization for multi-view 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16688–16698, 2024. 1

[40] 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. 3

[41] 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, 5, 6, 7

[42] 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. 2, 6, 7

[43] 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. 3

[44] Joni Pajarinen, Jens Lundell, and Ville Kyrki. Pomdp planning under object composition uncertainty: Application to robotic manipulation. *IEEE Transactions on Robotics*, 39 (1):41–56, 2022. 5

[45] 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. 3, 6

[46] 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. 1, 2, 3, 5, 6

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

[48] 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. 3, 6

[49] 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. 3, 6

[50] Tianqi Qie, Weida Wang, Chao Yang, Ying Li, Yuhang Zhang, Wenjie Liu, and Changle Xiang. An improved model

predictive control-based trajectory planning method for automated driving vehicles under uncertainty environments. *IEEE Transactions on Intelligent Transportation Systems*, 24 (4):3999–4015, 2022. 5

[51] 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

[52] 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. 3, 6

[53] Changyong Shu, Jiajun Deng, Fisher Yu, and Yifan Liu. 3dppe: 3d point positional encoding for transformer-based multi-camera 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3580–3589, 2023. 1

[54] 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. 5

[55] Sanbao Su, Yiming Li, Sihong He, Songyang Han, Chen Feng, Caiwen Ding, and Fei Miao. Uncertainty quantification of collaborative detection for self-driving. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5588–5594. IEEE, 2023. 5

[56] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 5

[57] Xiaolin Tang, Guichuan Zhong, Shen Li, Kai Yang, Keqi Shu, Dongpu Cao, and Xianke Lin. Uncertainty-aware decision-making for autonomous driving at uncontrolled intersections. *IEEE Transactions on Intelligent Transportation Systems*, 24(9):9725–9735, 2023. 5

[58] 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. 3

[59] Li Wang, Li Zhang, Yi Zhu, Zhi Zhang, Tong He, Mu Li, and Xiangyang Xue. Progressive coordinate transforms for monocular 3d object detection. *Advances in Neural Information Processing Systems*, 34:13364–13377, 2021. 6

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

[61] 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. 3

[62] 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

[63] 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. 3, 6

[64] 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, 3, 6

[65] 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. 1

[66] 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. 3

[67] 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. 3

[68] 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, 3, 6, 7

[69] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3289–3298, 2021. 3

[70] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 3, 5

[71] 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. 3, 6

[72] 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. 3, 7