Modeling and Learning Multiple Hypotheses for Monocular 3D Object Detection
Supplementary Material

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

A. Overview

- Appendix B: Ill-posedness of monocular 3D object detection.
- Appendix C: More analysis of single-point prediction.
- Appendix D: More analysis of hypothesis filtering.
- Appendix E: More details of base detector and implementation.
- Appendix F: More analysis of efficiency compared to the base detector.
- Appendix G: More experimental results and analysis.
- Appendix H: Qualitative results on the KITTI and Waymo datasets.
- Appendix I: Limitation.

B. Ill-Posedness of Monocular 3D Object Detection

In monocular 3D object detection, multiple plausible 3D bounding boxes can explain the same 2D observation of an object, even after eliminating infeasible solutions through geometric modeling. For example, several methods [9, 13] use approximate vehicle height, along with the assumption that vehicles rest on the ground plane, to infer depth information. However, inspired by previous works [6, 10, 12, 15, 22], our analysis below reveals that even a one-pixel change in the image space can correspond to a wide range of depth variations in the 3D world.

To better illustrate this phenomenon, we use a simple yet realistic camera setting, where the optical axis is parallel to the ground, to analyze the relationship between pixel positions in 2D image space and depth in 3D space. Based on the pinhole camera model, we can estimate the vehicle depth D from the focal length f, the vehicle height h observed in the image, and the known vehicle height H in the 3D space:

$$D = \frac{f \cdot H}{h} \tag{1}$$

If the observed 2D vehicle is 1 pixel taller, the estimated depth D' becomes:

$$D' = \frac{f \cdot H}{h+1} \tag{2}$$

The difference between these depth estimates is:

$$D - D' = \frac{f \cdot H}{h} - \frac{f \cdot H}{h+1} = \frac{D^2}{f \cdot H + D} \tag{3}$$

where we eliminate h and retain $f \cdot H$ in the final formula because h varies with the vehicle depth while f and H are constants for a fixed camera and a specific vehicle.

When the vehicle is far from the camera, the term $f \cdot H$ becomes negligible compared to D, making the depth difference approximately linear with D. This implies that even a one-pixel change in the image space can correspond to a wide range of depth variations in the 3D world.

C. More Analysis of Single-Point Prediction

When multiple plausible solutions exist for the same input, single-point prediction tends to regress toward the mean of these solutions. Below, we present a brief analysis of this phenomenon. A more comprehensive discussion can be found in the machine learning literature [1, 2, 17].

Let $p(\boldsymbol{t}, \boldsymbol{x})$ be the joint probability distribution of input variables $\boldsymbol{x}$ and target variables $\boldsymbol{t}$. In monocular 3D object detection, $\boldsymbol{x}$ could be a 2D object observed in a monocular image, $\boldsymbol{t}$ could be the 3D bounding box or position of the object, and the conditional probability distribution $p(\boldsymbol{t}|\boldsymbol{x})$ could be multimodal—meaning that the same 2D observation can correspond to multiple plausible 3D object configurations.

Let $f(\cdot; \boldsymbol{\theta})$ be a single-point predictor (e.g., a neural network) parameterized by $\boldsymbol{\theta}$. It takes an instance of $\boldsymbol{x}$ as input and outputs an estimate of $\boldsymbol{t}$. The training data consist of pairs of input and target values sampled from $p(\boldsymbol{t}, \boldsymbol{x})$: $\{(\boldsymbol{t}_n, \boldsymbol{x}_n) : n = 1, \ldots, N\}$, where N is the number of training samples.

Method	Car	Ped.	Cyc.
RoI	23.81 / 17.00 / 14.17	12.02 / 8.97 / 6.73	8.52 / 4.40 / 4.01
Central Window	23.68 / 17.04 / 14.12	11.39 / 9.09 / 6.82	8.06 / 4.22 / 3.84
Left Window	23.73 / 16.27 / 13.96	11.24 / 8.20 / 6.34	7.97 / 4.14 / 3.69
Top-right Window	22.54 / 15.87 / 13.61	10.39 / 7.50 / 6.02	9.05 / 4.61 / 4.12

Table 1. **Comparison of predictions from different RoI regions** on KITTI Validation Set. Performance is reported in the format: Easy/Moderate/Hard.

A common choice for the learning objective in regression problems is the ℓ_2-norm loss:

$$\frac{1}{2N}\sum_{n=1}^{N}\|f(\boldsymbol{x}_n;\boldsymbol{\theta})-\boldsymbol{t}_n\|_2^2 \tag{4}$$

In the limit of infinite training data, the ℓ_2-norm loss becomes:

$$\mathcal{L}=\frac{1}{2}\int\int p(\boldsymbol{t},\boldsymbol{x})\|f(\boldsymbol{x};\boldsymbol{\theta})-\boldsymbol{t}\|_2^2\,d\boldsymbol{t}\,d\boldsymbol{x} \tag{5}$$

Taking the derivative of $\mathcal{L}$ with respect to $f(\boldsymbol{x};\boldsymbol{\theta})$ and setting it to zero, we obtain:

$$\frac{\delta\mathcal{L}}{\delta f(\boldsymbol{x};\boldsymbol{\theta})}=\int p(\boldsymbol{t},\boldsymbol{x})(f(\boldsymbol{x};\boldsymbol{\theta})-\boldsymbol{t})\,d\boldsymbol{t}=0 \tag{6}$$

Rearranging this equation, the optimal single-point predictor is given by:

$$f^*(\boldsymbol{x};\boldsymbol{\theta})=\int p(\boldsymbol{t}|\boldsymbol{x})\boldsymbol{t}\,d\boldsymbol{t} \tag{7}$$

which corresponds to the conditional mean of $\boldsymbol{t}$ given $\boldsymbol{x}$.

If we replace the ℓ_2-norm loss with the ℓ_1-norm loss, the optimal single-point predictor instead corresponds to the conditional median of $\boldsymbol{t}$ given $\boldsymbol{x}$. Notably, both the conditional mean and conditional median may not correspond to any mode of $p(\boldsymbol{t}|\boldsymbol{x})$, meaning that the predicted value may not align with plausible solutions when the distribution is multimodal. Tab. 1 compares AP from single-point predictions using the RoI or individual windows, showing limited performance variation among them.

D. More Analysis of Hypothesis Filtering

For highly uncertain objects, the precision of all hypotheses is low, *e.g.*, around 20% precision in the confidence interval $[0.1, 0.2]$. In this case, retaining multiple hypotheses with similar precision levels will increase the likelihood of including accurate 3D bounding boxes without significantly compromising precision. We first present a simple analysis below and then show some empirical results.

Consider K hypotheses $\{\mathcal{H}_k : k = 1,\ldots,K\}$ after filtering. Suppose each hypothesis $\mathcal{H}_k$ has a probability p of having sufficient overlap with the ground-truth 3D bounding box. For single-point prediction, the expected number of true positives is p regardless of the chosen hypothesis. Hence, the precision and recall are $P_{\text{single}} = p$ and $R_{\text{single}} = p$, respectively. For multi-hypothesis prediction, the expected number of true positives is $1-(1-p)^K$ with K hypotheses. Therefore, the precision and recall are $P_{\text{multi}} = \frac{1-(1-p)^K}{K}$ and $R_{\text{multi}} = 1-(1-p)^K$, respectively.

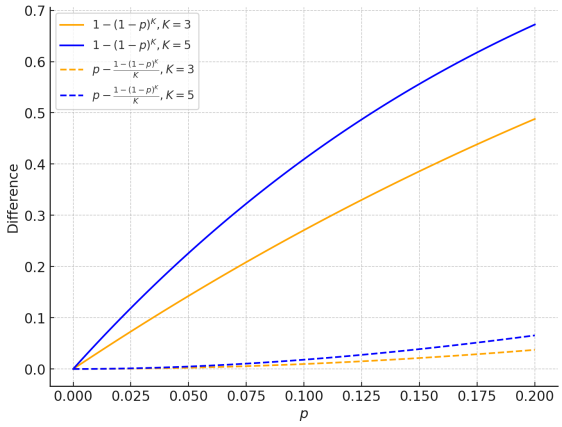
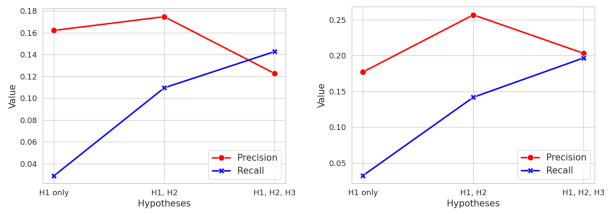


Figure 1. **Effect of multiple hypotheses on recall and precision.** For highly uncertain objects, retaining multiple (K) hypotheses with similar precision levels p will increase the recall (solid lines) without significantly compromising precision (dashed lines).

Obviously, multi-hypothesis prediction yields a recall improvement compared to single-point prediction $R_{\text{multi}} > R_{\text{single}}$: $1-(1-p)^K > p$. This inequality holds because $1-p > (1-p)^K$ for $p \in (0,1)$. The plot of $1-p-(1-p)^K$ in Fig. 1 also confirms this claim. To characterize the precision difference, we need to compare $P_{\text{single}} = p$ and $P_{\text{multi}} = \frac{1-(1-p)^K}{K}$. When p is close to 0, based on the first-order Taylor expansion: $(1-p)^K \approx 1-Kp$, which means P_{multi} will be roughly p. This aligns with the plot of $p-\frac{1-(1-p)^K}{K}$ shown in Fig. 1.

To empirically validate the analysis, we conduct experiments on the KITTI dataset, focusing on the performance of uncertain objects under varying confidence ranges. As can be seen in Fig. 2, we observe the same trend by evaluating the precision and recall changes using three hypotheses. Compared to the single-point estimation approach, where only one hypothesis is used, employing all three hypotheses demonstrates a slight decrease in precision. However, this minor precision drop was outweighed by a substantial improvement in recall. From this perspective, multi-hypothesis filtering also plays a role in enhancing the likelihood of detecting accurate 3D bounding boxes, particularly for uncertain objects. As a result, retaining multiple hypotheses demonstrates a practical trade-off, enhancing recall without significantly degrading precision.



(a) (b)

Figure 2. **The trend of recall and precision in the confidence ranges** (a) [0, 0.1] and (b) [0.1, 0.2] on the KITTI dataset.

E. More Details of Base Detector and Implementation.

We adopt a common architecture of recent state-of-the-art methods [3, 7, 9, 11, 19, 21] as a generic base detector. It consists of a feature backbone, 2D detection heads, and 3D detection heads.

Backbone and 2D Detection Heads. Taking as input an image $\boldsymbol{I} \in \mathbb{R}^{W\times H\times 3}$, where W and H are the image width and height, the backbone network (*i.e.*, DLA34 [20]) produces a feature map $\boldsymbol{F} \in \mathbb{R}^{W/4\times H/4\times C}$, where C is the feature dimension. The feature map $\boldsymbol{F}$ is then fed into three separate 2D detection heads, each predicting specific outputs: a 2D heatmap $\boldsymbol{H} \in \mathbb{R}^{W/4\times H/4\times B}$ where B is the number of categories, a 2D size map $\boldsymbol{S}^{\text{2D}} \in \mathbb{R}^{W/4\times H/4\times 2}$ for bounding box dimensions, and a 2D offset map $\boldsymbol{O}^{\text{2D}} \in \mathbb{R}^{W/4\times H/4\times 2}$ for positional adjustment. By combining these predictions, the 2D bounding boxes can be obtained. Finally, RoI-Align is applied to extract object features $\boldsymbol{F}^{\text{roi}} \in \mathbb{R}^{N\times M\times M\times C}$, where N is the number of RoIs, and $M \times M$ denotes the RoI region size.

3D Detection Heads. The RoI features $\boldsymbol{F}^{\text{roi}}$ are passed into 3D detection heads to estimate 3D bounding boxes. These heads regress 3D sizes $\boldsymbol{S}^{\text{roi}} \in \mathbb{R}^{N\times 3}$, orientations $\boldsymbol{\Theta}^{\text{roi}} \in \mathbb{R}^{N\times k\times 2}$ (using a multi-bin approach [9], with k bins), and offsets for the 3D center projection $\boldsymbol{O}^{\text{3D}} \in \mathbb{R}^{N\times 2}$. Additionally, two heads respectively predict depth maps $\boldsymbol{D}^{\text{pixel}} \in \mathbb{R}^{N\times M\times M}$, and uncertainty maps $\boldsymbol{U}^{\text{roi}} \in \mathbb{R}^{N\times M\times M}$. The uncertainty $\boldsymbol{U}^{\text{roi}}$ is converted to confidence as $\boldsymbol{C}^{\text{roi}} = \sigma(-\boldsymbol{U}^{\text{roi}})$, where $\boldsymbol{C}^{\text{roi}} \in \mathbb{R}^{N\times M\times M}$.

MonoMH builds upon this base detector and introduces three key innovations described in the paper: multi-hypothesis prediction, multi-hypothesis learning, and hypothesis filtering. We keep the depth estimate and supervision in the base detector for better performance.

Architectural Setup. The backbone model is initialized with ImageNet weights, while the detection heads are initialized using the Xavier algorithm [4]. All feature dimensions are set to 256, with each head's output dimensions varying based on the type of prediction (*e.g.*, offset, size, and orientation). For both datasets, we predict three foreground categories, excluding the 'Sign' class in the Waymo dataset [5, 14]. In RoI Division, the uncertainty estimate is scaled by -0.01 before being fed to a sigmoid function for confidence prediction. This ensures that high uncertainty results in low confidence and avoids being over-sensitive to noisy uncertainty estimates.

KITTI	GPU memory (GB)		Time (ms/image)		Model
	Train.	Infer.	Train.	Infer.	Size (M)
Base detector	2.55	1.66	45.2	25.9	20.52
Ours	2.56	1.69	48.5	26.6	20.68

Table 2. **Efficiency analysis.** 'Train.' and 'Infer.' indicate training and inference, respectively.

Data Augmentation. During training, we apply horizontal flipping with a probability of 0.5 and scale augmentation with a factor of 0.4 [5, 7, 9]. For the KITTI dataset, we additionally employ Mixup3D [7] with a mix proportion of 0.5 to address its smaller training set size and prevent overfitting.

Pre-processing. Following the conventional settings [5, 7, 11], KITTI images with a resolution of [370, 1242] and Waymo images with a resolution of [1280, 1920] are resized to [384, 1280] and [512, 768], respectively.

Training. For Waymo, we follow the settings in [5, 10], including boxes with more than 100 LiDAR points for the vehicle class and more than 50 LiDAR points for the cyclist and pedestrian classes. Weight decay is set to 0.00001, and the learning rate decays by a factor of 0.1 at the 120th and 160th epochs for the validation set, and at the 240th, 360th, and 480th epochs for the test set. All loss terms are assigned equal weights. In our configuration settings, training the KITTI split set or the Waymo dataset takes approximately 18 hours, while training the full KITTI training set for test set evaluation requires about 2 days.

Inference. We limit the maximum number of objects per image to 50. For filtering, we discard 2D boxes with category confidence lower than 0.2 for KITTI and 0.1 for Waymo.

F. More Analysis of Efficiency Compared to the Base Detector

Tab. 2 compares GPU memory usage, training time, inference time, and model size between our method and the base detector. All evaluations are conducted on a single V100 GPU with a batch size of 1. MonoMH introduces minimal computational overhead in terms of the network's training, inference, and model size compared to the base detector.

G. More Experimental Results and Analysis

Additional Analysis of Multi-hypothesis Prediction. Fig. 3 further verifies the diversity of hypotheses produced by

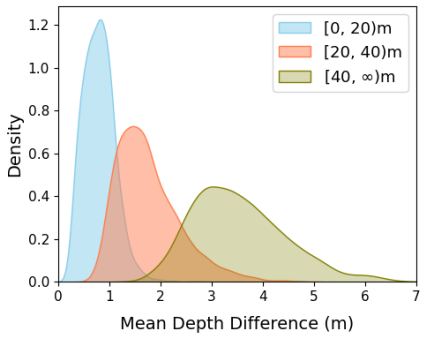


Figure 3. **Distribution of mean depth differences between hypotheses generated by RoI Division within an RoI** across three depth ranges. RoI Division yields broader depth hypotheses for farther (more ambiguous) objects.

Method	Ped. (E/M/H)	Cyc. (E/M/H)
Baseline	12.02/8.97/6.74	8.52/4.40/4.01
RoI Div., MH loss	14.84/10.96/8.48	9.65/4.80/4.35
RoI Div., MH loss, filtering	**15.67/11.78/9.34**	**10.55/5.60/5.08**

Table 3. **Comparison of main contributions for the *Pedestrian* and *Cyclist* categories.** 'RoI Div.' means RoI Division, 'MH loss' means soft BoM loss, 'filtering' means hypothesis filtering. Performance is reported in the format: Easy/Moderate/Hard. Detailed ablation experiments for the *Car* category and alternative designs are provided in the manuscript.

RoI Division. As distance increases, depth diversity increases. This indicates that RoI Division generates broader depth hypotheses to reflect depth ambiguity that far objects exhibit.

Main Ablation Results on Other Categories. Tab. 3 shows main ablation results on 'Pedestrian' and 'Cyclist'. The soft BoM loss is used as the multi-hypothesis (MH) loss, and the most confident hypothesis is retained from RoI Division when hypothesis filtering is not applied. Note that applying RoI Division without any multi-hypothesis loss is meaningless. The results reconfirm the effectiveness of the main contributions of MonoMH.

Comparison of Hypotheses Processing Methods on Other Categories. To further 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, on 'Pedestrian' and 'Cyclist'. Our filtering strategy yields the best results across all difficulties over the other methods, reconfirming that keeping a small set of plausible hypotheses under uncertainty is most effective.

Impact of Hyper-parameters. Tab. 5 shows the performance of MonoMH under varying window sizes and numbers of hypotheses. The results demonstrate that MonoMH is highly robust to these hyper-parameters and consistently outperforms the baseline by a substantial margin. Since the 4×4 window with nine hypotheses provides reliable performance, we adopt this setting for all other experiments without category-specific tuning.

Method	Ped.	Cyc.
Mean	14.69/10.60/8.64	8.21/4.29/3.80
Best	14.84/10.96/8.48	9.65/4.80/4.35
Filtering	**15.67/11.75/9.34**	**10.55/5.60/5.08**

Table 4. **Comparison of hypotheses processing methods for the *Pedestrian* and *Cyclist* categories.** Performance is reported in the format: Easy/Moderate/Hard.

w size	# h	Car	Ped.	Cyc.
2×2	5	27.61/19.86/16.93	14.75/10.82/8.63	9.45/4.92/4.36
	9	27.89/20.32/17.43	14.85/10.94/8.64	10.00/5.26/4.74
	16	27.41/20.00/17.05	14.55/10.62/8.36	9.11/4.86/4.36
3×3	5	27.61/19.86/16.93	14.75/10.82/8.63	9.45/4.92/4.36
	9	28.13/20.91/18.02	14.99/11.14/8.85	10.06/5.38/4.85
	16	27.67/20.20/17.33	14.45/10.47/8.27	8.78/4.57/4.20
4×4	5	28.63/21.15/18.20	15.21/11.27/8.90	10.53/5.49/5.05
	9	**28.94/21.53/18.43**	**15.67/11.78/9.34**	10.55/5.60/5.08
	16	28.11/20.63/17.67	15.11/11.18/8.83	9.36/4.94/4.51
5×5	5	28.67/20.97/18.11	15.02/11.39/8.96	**10.91**/5.70/**5.30**
	9	28.14/20.46/17.48	15.09/11.34/8.98	10.76/**5.79**/5.29

Table 5. **Impact of the window size and the number of hypotheses.** Performance is reported in the format: Easy/Moderate/Hard.

Method	N=1,000			N=5,000			N=9,000		
	TP (↑)	FP (↓)	FN (↓)	TP (↑)	FP (↓)	FN (↓)	TP (↑)	FP (↓)	FN (↓)
MonoLSS	516	484	2,331	**1,201**	**3,799**	**1,646**	1,223	**7,212**	1,624
+ Ours	**533**	**467**	**2,314**	1,187	3,813	1,660	**1,374**	7,626	**1,473**
MonoDETR	490	510	2,357	**1,316**	**3,684**	**1,531**	1,351	**6,407**	1,496
+ Ours	**516**	**484**	**2,331**	1,216	3,784	1,631	**1,509**	7,491	**1,338**
Base.	513	487	2,334	1,182	3,818	1,665	1,204	**7,130**	1,643
MonoMH	**553**	**447**	**2,294**	**1,195**	**3,805**	**1,652**	**1,357**	7,643	**1,490**

Table 6. **TP/FP/FN analysis for *Car* category at the Easy difficulty.** For each method, we keep the top-N detections on validation data and report TP/FP/FN for single prediction baselines and their multi-hypothesis counterpart ('+ Ours'). The total number of ground truth detections on the validation data is 2,847.

TP/FP/FN Analysis. Tab. 6 reports counts on KITTI val for the 'Car' category at 'Easy' difficulty with fixed detection budgets. For each method, we keep the global top-N detections by confidence to compare the number of TP/FP/FN. For the small number of detections, such as 1,000 and 5,000, our multi-hypothesis variants are better or comparable to their single-prediction counterparts. As the detection number increases (*i.e.*, 9,000), they consistently convert more detections into TP and reduce FN, indicating that keeping multiple plausible modes helps recover missed objects. Importantly, these gains come with little FP inflation for MonoLSS and the Base detector (often slightly lower FP), while MonoDETR shows a modest FP rise yet still improves TP and FN. Overall, the trend indicates that MonoMH improves recall (safety) without sig-

nificantly sacrificing precision for strong baselines.

H. Qualitative Results

Qualitative Results on KITTI. The qualitative results in Fig. 4 demonstrate the accuracy of our MonoMH predictions (pink bounding boxes) in capturing the 3D positions and dimensions of objects. MonoMH predicts a single 3D bounding box that aligns well with the ground truth (green boxes) in the Bird's-Eye-View (BEV) space for certain objects that are nearby or not occluded. In contrast, for uncertain objects, such as those heavily occluded or located at a far distance where accurate prediction is challenging, MonoMH predicts multiple bounding boxes, effectively increasing the likelihood of capturing the correct bounding box. The visualizations highlight MonoMH's robustness in providing reliable 3D localization for handling diverse environments.

Moreover, Fig. 5 illustrates MonoMH in dense urban scenes, including all class categories. For confident objects (typically nearer and unoccluded), MonoMH outputs a single 3D box that aligns well with ground truth in BEV. For uncertain objects, it retains a small set of spatially clustered hypotheses, increasing the chance of capturing the correct box. Overall, the visualizations highlight that MonoMH provides reliable 3D localization while maintaining practicality in scenes with substantial clutter and occlusion.

Qualitative Results on Waymo. Fig. 6 illustrates the effectiveness of our proposed MonoMH compared to the baseline on the Waymo dataset. The pink bounding boxes (ours) demonstrate improved alignment with the bright green ground truth boxes (in the BEV space), especially in challenging cases involving objects at greater distances or severely occluded, as highlighted in the magnified regions. In contrast, the blue bounding boxes (baseline) often fail to accurately capture the object's 3D position and size, exhibiting higher deviations. These results emphasize the capability of our multi-hypothesis framework to better handle depth ambiguity and generate more precise 3D bounding boxes, even in complex scenarios.

I. Limitation

The uncertainty estimates in existing methods [8, 9, 11, 16, 18] often do not exhibit a perfect linear relationship with depth accuracy. We found that MonoMH still faces this common limitation. Improving the uncertainty estimates could benefit both MonoMH and other state-of-the-art methods, which we plan to investigate in future work.

Figure 4. **Qualitative results on KITTI.** Each row shows the image with projected 3D boxes (left) and the corresponding BEV (right). MonoMH predictions align closely with the Ground Truth in the BEV space. Particularly for uncertain objects, such as those occluded or distant, MonoMH generates multiple bounding boxes, increasing the likelihood of accurate prediction.

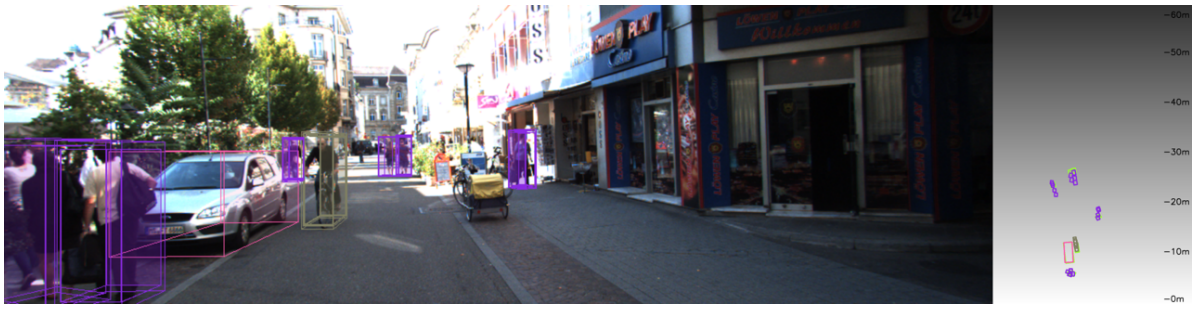


Figure 5. **Qualitative result on a crowded scene in KITTI.** The image with projected 3D boxes (left) and the corresponding BEV (right) are shown. To validate our method in a crowded scene, we illustrate all classes: car (pink), pedestrian (purple), and cyclist (olive). MonoMH yields a small, spatially clustered set of hypotheses only for uncertain objects while keeping single top predictions for confident ones, illustrating practicality under heavy occlusion and clutter.

Figure 6. **Qualitative results on Waymo.** Each row shows the image with projected 3D boxes (left) and the corresponding BEV (right). MonoMH can capture more precise 3D bounding boxes that align closely with Ground Truth in BEV space, outperforming the baseline. Magnified regions emphasize MonoMH's ability to address depth ambiguity from distant or occluded objects, resulting in accurate 3D localization.

References

[1] Christopher M Bishop. Mixture density networks. 1994. 1

[2] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006. 1

[3] 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

[4] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 3

[5] 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. 3

[6] 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. 1

[7] 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. 3

[8] Lijie Liu, Jiwen Lu, Chunjing Xu, Qi Tian, and Jie Zhou. Deep fitting degree scoring network for monocular 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1057–1066, 2019. 5

[9] 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, 3, 5

[10] 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, 3

[11] 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. 3, 5

[12] Liang Peng, Senbo Yan, Chenxi Huang, Xiaofei He, and Deng Cai. Digging into output representation for monocular 3d object detection, 2022. 1

[13] 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

[14] 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

[15] 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. 1

[16] 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

[17] Stephen M Stigler. Regression towards the mean, historically considered. *Statistical methods in medical research*, 6(2): 103–114, 1997. 1

[18] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*, pages 1475–1485. PMLR, 2022. 5

[19] 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. 3

[20] 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

[21] 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

[22] 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. 1