# Monocular 3D Object Detection with LiDAR Guided Semi Supervised Active Learning (Supplementary Material)

Aral Hekimoglu
Technical University Munich
Munich, Germany
aral.hekimoglu@tum.de

Michael Schmidt
BMW Group
Munich, Germany
michael.se.schmidt@bmw.de

Alvaro Marcos-Ramiro
BMW Group
Munich, Germany
alvaro.marcos-ramiro@bmw.de

## 1. Derivation of Eq. (2)

For simplicity, we use $f$ as $f_s$ and rewrite Eq. (1) with the MSE loss function as follows:

$$E[L] = \iint (f(x) - y)^2 dxdy \tag{16}$$

The optimization goal is to find an f(x) that minimizes $E[L]$. If we assume a completely flexible function f(x), we can do this formally by taking the derivative to give

$$\frac{\delta E[L]}{\delta f(x)} = 2 \int (f(x) - y)dy = 0 \tag{17}$$

We define $h(x)$ as the optimal function that satisfies this equation. Adding and subtracting $h(x)$ to Eq. (16) gives

$$E[L] = \iint (f(x) - h(x) + h(x) - y)^2 dxdy \tag{18}$$

$$\begin{aligned} E[L] = &\iint (f(x) - h(x))^2 + (h(x) - y)^2 dxdy \\ &+ 2 \iint (f(x) - h(x))(h(x) - y)dxdy \end{aligned} \tag{19}$$

where the final term is zero because for the optimal network $\int (h(x) - y)dy = 0$. Therefore, for a single point $x$, we can write the loss function as Eq. (2).

## 2. Derivation of Eq. (12)

We start from the loss definition in Eq. (2). For simplicity of notation, we use $f(x)$ instead of $f_s(x; \theta)$ and define expectation over the model parameters $E_\theta[f_s(x; \theta)]$ as $\mu$. Adding and subtracting $\mu$ to Eq. (2) and expanding results in the following equation:

$$\begin{aligned} L(x) = &(f(x) - \mu)^2 + (\mu - h(x))^2 \\ &+ 2 * (f(x) - \mu)(\mu - h(x)) \\ &+ (h(x) - y)^2 \end{aligned} \tag{20}$$

For reasons mentioned in the paper, we take the expectation of the loss over the model parameters $E_\theta$.

$$\begin{aligned} E_\theta[L(x)] = &E_\theta[(f(x) - \mu)^2] + (\mu - h(x))^2 \\ &+ 2 * E_\theta[(f(x) - \mu)(\mu - h(x))] \\ &+ (h(x) - y)^2 \end{aligned} \tag{21}$$

Note that the third term disappears, as shown below:

$$\begin{aligned} &E_\theta[(f(x) - \mu)(\mu - h(x))] \\ &= E_\theta[f(x) * \mu - h(x) * f(x) - \mu^2 + \mu * h(x)] \\ &= \mu^2 - h(x) * \mu - \mu^2 + \mu * h(x) = 0 \end{aligned} \tag{22}$$

Then Eq. (21) is written as,

$$E_\theta[L(x)] = E_\theta[(f(x)-\mu)^2]+(\mu-h(x))^2+(h(x)-y)^2 \tag{23}$$

where $(h(x) - y)^2$ is the aleatoric uncertainty $u^{al}$. For the student model $f_s(x, \theta)$, Eq. (23) results in Eq. (12).

## 3. Statistics of Eq. (15)

We provide the statistical summary for the components in Eq. (15) from the main paper. $u_s^{tv}$ and $i_{ts}$ have a range from 0 to 3.49, a mean of 0.87, and a variance of 0.28. Similarly, $u_t^{al}$ has a range from 0.19 to 1.23, a mean of 0.25, and a variance of 0.16.

| $\lambda_U$ | Mod. | Easy | Hard |
|---|---|---|---|
| 0.1 | 24.48 | 34.51 | 21.37 |
| 0.2 | 26.83 | 36.33 | 24.04 |
| 0.5 | **27.92** | **36.86** | 26.03 |
| 0.7 | 27.37 | 35.28 | **26.51** |
| 1.0 | 25.77 | 35.02 | 24.69 |
| 2.0 | 24.38 | 34.94 | 22.34 |

Table 5. The effect of hyperparameter $\lambda_u$ on the BEV AP performance of the monocular detector on KITTI *val*

## 4. Weight of unlabeled samples $\lambda_U$

Tab. 5 reports the experimental results on finetuning the hyperparameter for the unsupervised loss weight $\lambda_U$. We discovered that having a small or large $\lambda_U$ drops the performance. Therefore, finding a balanced ratio between supervised and unsupervised loss is important. Based on our findings in the ablation study, we set $\lambda_U = 0.5$ for all trainings in the main paper.

## 5. Effect of aleatoric head on LiDAR detector

We investigate the impact of aleatoric uncertainty on the performance of the PV-RCNN LiDAR detector, as presented in Tab. 6. Our findings indicate that aleatoric uncertainty has a negligible effect on the detector's performance, as evidenced by the slight decrease of only 0.15 and 0.25 in the moderate and easy AP scores, respectively, and a modest increase of 0.09 in the hard AP score.

| Aleatoric | Mod. | Easy | Hard |
|---|---|---|---|
| X | 82.58 | 89.95 | 77.32 |
| ✓ | 82.43 | 89.70 | 77.41 |

Table 6. Comparison of aleatoric uncertainty head on the BEV AP performance of the LiDAR detector on KITTI *val*.

## 6. 3D detection results on KITTI

We provide the comparison of 3D results in Tab. 7 on the KITTI *test* set. Among the methods that use semi-supervised LiDAR guidance, our approach reaches +1.06 and +2.83 BEV AP than the SOTA LPCG and MonoDistill, respectively. Considering the performance weighted by the number of samples in each case, MonoLiG has a higher overall AP of 19.75 compared to 19.15 of LPCG.

## 7. Effect of extra data on KITTI classes

We present the effect of the number of extra samples used for semi-supervised learning on the performance of the MonoFlex detector [54] in Tab. 8. We observe that as more data is trained with our semi-supervised strategy, performance for the *Car* and *Cyclist* classes increases. *Pedestrian* performance is affected less by the semi-supervised training and even decreases for certain experiments, and we attribute this to LiDAR having a lower performance on the *Pedestrian* class compared to the other two classes. This solidifies our conclusion in the main paper that as the performance of the teacher model increases, we expect better performance on the student model.

| Approaches | Extra | Mod. | Easy | Hard |
|---|---|---|---|---|
| M3D-RPN | - | 9.71 | 14.76 | 7.42 |
| MonoRUn | - | 12.30 | 19.65 | 10.58 |
| DDMP-3D | KD | 12.78 | 19.71 | 9.80 |
| PCT | KD | 13.37 | 21.00 | 11.31 |
| MonoFlex | - | 13.89 | 19.94 | 12.07 |
| MonoDTR | - | 15.39 | 21.99 | 12.73 |
| DID-M3D | - | 16.29 | 24.40 | 13.75 |
| DD3D | DDAD | 16.87 | 23.19 | 14.36 |
| MonoDDE | - | 17.14 | 24.93 | 15.10 |
| MonoDistill | - | 16.03 | 22.97 | 13.60 |
| LPCG | KD | 17.80 | **25.56** | 15.38 |
| MonoLiG | KD | **18.86** | 24.90 | **16.79** |

Table 7. Comparison of 3D detection results on KITTI *test* for monocular detectors. Note that both LPCG and MonoLiG use MonoFlex as the base detector. MonoDistill, LPCG, and MonoLiG are semi-supervised methods using additional information from LiDAR during training. KD and DDAD represent the extra datasets, KITTI-depth and DDAD15M, respectively.

## 8. Exact values from AL figures

Due to the limited space in the main paper, we present our AL comparisons as plots. Tab. 9, Tab. 10, Tab. 11 provides the exact metric values for Figures 4a, 4b, 4c from the main paper. The mean and variances of three experiments trained with different random initializations are presented.

## 9. Qualitative results on KITTI

We present predictions on KITTI dataset from the base DD3D detector [28] and DD3D trained with MonoLiG in Fig. 6. We show some of the best cases along with the failure cases. Our method localizes the *Car* class better in the BEV space, and our predictions are closer to the ground-truth boxes compared to the base detector, but for the *Pedestrian* and *Cyclist* class our approach has more false negatives, which the base detector detects but our detector fails.

| # of Extra | Vehicle | | | Pedestrian | | | Cyclist | | |
|---|---|---|---|---|---|---|---|---|---|
| | Mod. | Easy | Hard | Mod | Easy | Hard | Mod. | Easy | Hard |
| 0% | 14.96 | 20.09 | 13.62 | 6.39 | 8.76 | 4.75 | 3.30 | 4.49 | 2.58 |
| 10% | 15.24 | 20.46 | 13.72 | 6.24 | 8.43 | 4.41 | 3.14 | 4.38 | 2.44 |
| 20% | 15.82 | 20.41 | 14.42 | **6.42** | 8.34 | **4.87** | 3.32 | 4.67 | 2.50 |
| 30% | 15.63 | 20.75 | 14.46 | 6.36 | **8.97** | 4.26 | 3.60 | 5.10 | 2.89 |
| 40% | 16.55 | 21.39 | 14.94 | 6.02 | 8.79 | 4.12 | 3.85 | 5.20 | 2.74 |
| 50% | **16.90** | **22.07** | **15.27** | 6.25 | 8.88 | 4.38 | **4.10** | **5.64** | **2.96** |

Table 8. Effect of number of unlabeled samples in the 3D AP performance. We use KITTI-depth dataset explained in the main paper and randomly select a portion of it for each experiment.

| | 30 | 40 | 50 | 60 | 70 | 80 | 90 |
|---|---|---|---|---|---|---|---|
| Random | 8.04±0.23 | 9.36±0.16 | 10.45±0.23 | 11.32±0.15 | 13.10±0.10 | 13.71±0.13 | 13.73±0.25 |
| Entropy | 8.04±0.23 | 9.10±0.30 | 10.18±0.10 | 11.43±0.25 | 13.23±0.18 | 13.90±0.12 | 13.85±0.25 |
| Core-Set | 8.04±0.23 | 9.76±0.16 | 10.62±0.10 | 11.75±0.16 | 13.76±0.27 | 13.90±0.10 | 13.80±0.25 |
| LL4AL | 8.04±0.23 | 9.83±0.22 | 11.47±0.11 | 11.53±0.13 | 13.87±0.15 | 14.05±0.27 | 14.34±0.17 |
| CDAL | 8.04±0.23 | 10.72±0.24 | 11.90±0.17 | 12.18±0.13 | 14.19±0.26 | 14.20±0.19 | 14.24±0.19 |
| MonoLiG | 8.04±0.23 | 10.85±0.17 | 12.40±0.16 | 13.13±0.19 | 14.78±0.11 | 14.97±0.13 | 15.14±0.20 |

Table 9. Comparison with SOTA AL methods with semi-supervised training on KITTI *val* (Fig. 4a).

| | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 |
|---|---|---|---|---|---|---|---|---|
| Random | 3.50±0.10 | 4.16±0.07 | 4.36±0.13 | 4.50±0.05 | 4.60±0.13 | 4.80±0.07 | 5.06±0.06 | 5.30±0.08 |
| Entropy | 3.50±0.10 | 4.30±0.08 | 4.32±0.12 | 4.62±0.12 | 4.70±0.12 | 4.94±0.11 | 5.22±0.12 | 5.36±0.07 |
| LL4AL | 3.50±0.10 | 4.34±0.14 | 4.41±0.08 | 4.74±0.07 | 4.93±0.10 | 5.18±0.12 | 5.32±0.10 | 5.55±0.15 |
| CDAL | 3.50±0.10 | 4.36±0.13 | 4.44±0.14 | 4.86±0.11 | 5.09±0.10 | 5.27±0.06 | 5.47±0.10 | 5.63±0.14 |
| Core-Set | 3.50±0.10 | 4.38±0.15 | 4.50±0.15 | 4.96±0.06 | 5.15±0.11 | 5.23±0.07 | 5.39±0.11 | 5.54±0.14 |
| MonoLiG | 3.50±0.10 | 4.47±0.05 | 4.63±0.11 | 5.14±0.14 | 5.33±0.07 | 5.44±0.11 | 5.68±0.06 | 5.88±0.09 |

Table 10. Comparison with SOTA AL methods with semi-supervised training on Waymo *val*. (Fig. 4b).

| | 30 | 40 | 50 | 60 | 70 | 80 | 90 |
|---|---|---|---|---|---|---|---|
| Random | 7.81±0.18 | 9.17±0.26 | 10.10±0.21 | 10.39±0.27 | 12.06±0.21 | 12.50±0.27 | 13.56±0.16 |
| Entropy | 7.81±0.18 | 8.59±0.19 | 10.00±0.17 | 10.65±0.30 | 12.47±0.21 | 12.93±0.15 | 13.64±0.11 |
| Core-Set | 7.81±0.18 | 9.33±0.11 | 10.58±0.21 | 10.73±0.25 | 12.74±0.20 | 13.62±0.14 | 13.71±0.21 |
| LL4AL | 7.81±0.18 | 9.36±0.21 | 11.00±0.26 | 11.28±0.15 | 12.70±0.13 | 13.52±0.26 | 13.69±0.10 |
| CDAL | 7.81±0.18 | 10.23±0.26 | 11.38±0.13 | 12.12±0.23 | 13.27±0.12 | 13.80±0.17 | 14.04±0.22 |
| MonoLiG | 7.81±0.18 | 10.18±0.22 | 11.43±0.19 | 12.41±0.18 | 13.76±0.21 | 14.12±0.14 | 14.65±0.10 |

Table 11. Comparison with SOTA AL methods with supervised training on KITTI *val*. (Fig. 4c).

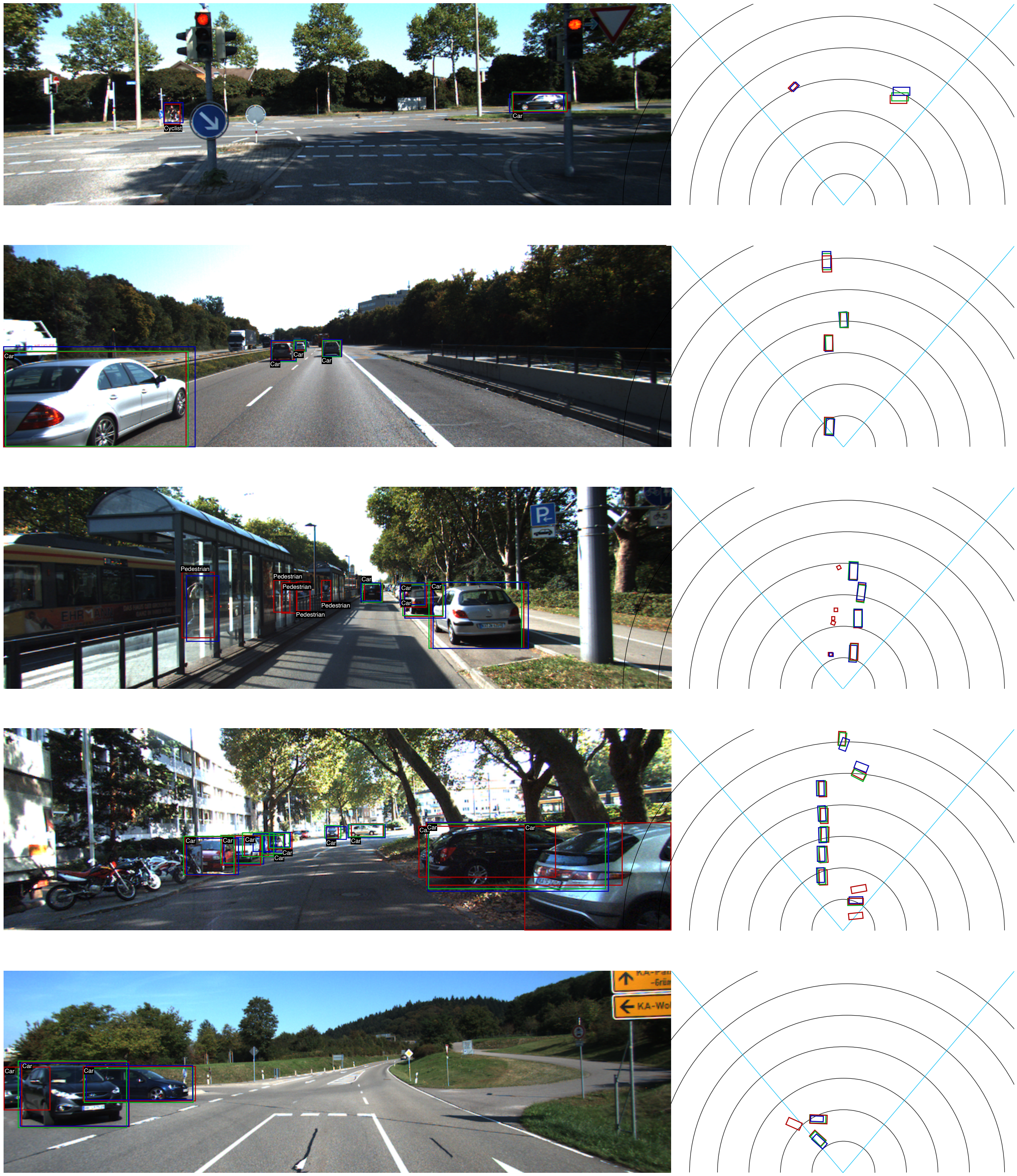

Figure 6. Qualitative comparison on KITTI. Each row shows the image and 3D predictions projected to the image on the left and BEV predictions on the right. Red, blue, and green indicate ground-truth, DD3D, and MonoLiG, respectively.