CHARM3R: Towards Unseen Camera Height Robust Monocular 3D Detector

Supplementary Material

Contents

A1. Additional Details and Proof

We now add more details and proofs which we could not put in the main paper because of the space constraints.

A1.1. Proof of Ground Depth Lemma 1

We reproduce the proof from [110] with our notations for the sake of completeness of this work.

Proof. We first rewrite the pinhole projection Eq. (1) as:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \mathbf{R}^{-1}(\mathbf{K}^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} z - \mathbf{T}). \tag{9}$$

We now represent the ray shooting from the camera optical center through each pixel as $\overrightarrow{r}(u,v,z)$. Using the matrix $\boldsymbol{A} = (a_{ij}) = \mathbf{R}^{-1}\mathbf{K}^{-1}$, and the vector $\boldsymbol{B} = (b_i) = -\mathbf{R}^{-1}\mathbf{T}$, we define the parametric ray as:

$$\overrightarrow{r}(u,v,z) : \begin{cases} X = (a_{11}u + a_{12}v + a_{13})z + b_1 \\ Y = (a_{21}u + a_{22}v + a_{23})z + b_2 \\ Z = (a_{31}u + a_{32}v + a_{33})z + b_3 \end{cases} \tag{10}$$

Moreover, the ground at a distance h can be described by a plane, which is determined by the point $(0, H, 0)$ in the plane and the normal vector $\overrightarrow{n} = (0, 1, 0)$:

$$\overrightarrow{r} \cdot \overrightarrow{n} = H. \tag{11}$$

Then, the ground depth is the intersection point between this ray and the ground plane. Combining Eqs. (10) and (11), the ground depth z of the pixel (u, v) is:

$$(a_{21}u + a_{22}v + a_{23})z + b_2 = H$$

$$\implies z = \frac{H - b_2}{a_{21}u + a_{22}v + a_{23}}. \tag{12}$$

□

A1.2. Proof of Lemma 2

We next derive Lemma 2 from Lemma 1 as follows.

Proof.

$$\boldsymbol{A} = (a_{ij}) = \mathbf{R}^{-1}\mathbf{K}^{-1} = \boldsymbol{I}^{-1} \begin{bmatrix} f & 0 & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{bmatrix}^{-1}$$
$$= \boldsymbol{I} \begin{bmatrix} \frac{1}{f} & 0 & \frac{-u_0}{f} \\ 0 & \frac{1}{f} & \frac{-v_0}{f} \\ 0 & 0 & 1 \end{bmatrix},$$

with rotation matrix $\mathbf{R}$ is identity $\boldsymbol{I}$ for forward cameras. So, $a_{21} = 0, a_{22} = \frac{1}{f}, a_{23} = \frac{-v_0}{f}$. Substituting a_{21}, a_{22}, a_{23} in Eq. (2), we get Eq. (3). □

A1.3. Pixel Shift with Ego Height Change.

We derive pixel shift with ego height change by backprojecting a pixel $\mathbf{p} = (u, v, z)$ to 3D, applying extrinsics change, and re-projecting to 2D. The new point $\mathbf{p}'$ after height change of ΔH is given by $\mathbf{p}' = \mathbf{K}[\mathbf{R}|\mathbf{t}]\mathbf{K}^{-1}\mathbf{p}$ with usual notations. With height change inducing translation $\mathbf{t} = [0, \Delta H, 0]^T$ and not changing rotations ($\mathbf{R} = \mathbf{I}$), $\mathbf{p}' = (u, v + \frac{f\Delta H}{z}, z)$.

A1.4. Extension to Camera Not Parallel to Ground

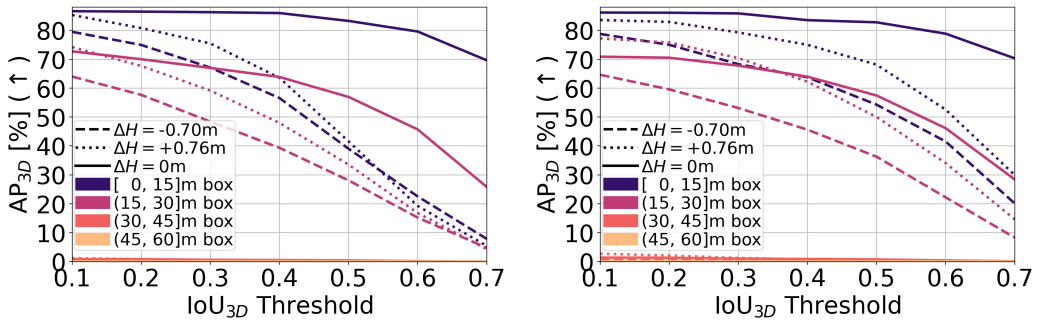
Following Sec. 3.3 of GEDepth [110], we use the camera pitch δ, and generalize Eq. (2) to obtain ground depth as

$$z = \frac{H - b_2\cos\delta - b_3\sin\delta}{[a_{21}u + a_{22}v + a_{23}]\cos\delta + [a_{31}u + a_{32}v + a_{33}]\sin\delta}$$
$$= \frac{H - b_2\cos\delta - b_3\sin\delta}{\frac{v - v_0}{f}\cos\delta + \sin\delta} \tag{13}$$

. Note that if camera pitch $\delta = 0$, this reduces to the usual form of Eq. (2) and Eq. (3) respectively. Also, Th. 1 has a more general form with the pitch value, and remains valid for majority of the pitch angle ranges.

A1.5. Extension to Not-flat Roads

For non-flat roads, we assume that the road is made of multiple flat 'pieces' of roads each with its own slope and we predict the slope of each pixel as in GEDepth [110]. To



(a) GUP Net (b) CHARM3R

Figure 7. **CARLA Val AP_{3D} at different depths and IoU_{3D} thresholds** with GUP Net. CHARM3R shows biggest gains on $IoU_{3D} > 0.3$ for $[0, 30]m$ boxes. Note that $30m$ to $60m$ curves are present, but their performance is near zero, making them hard to see.

predict slope $\hat{\delta}$ of each pixel, we first define a set of N discrete slopes: $\{\tau_i, i = 1, ..., N\}$. We compute each pixel slope by linearly combining the discrete slopes with the predicted probability distribution $\{\hat{p_i} \in [0, 1], \sum_i \hat{p_i} = 1\}$ over N slopes $\hat{\delta} = \sum_i \hat{p_i}\tau_i$. We train the network to minimize the total loss: $L_{\text{total}} = L_{\text{det}} + \lambda_{\text{slope}} L_{\text{slope}}(\delta, \hat{\delta})$, where L_{det} are the detection losses, and L_{slope} is the slope classification loss. We next substitute the predicted slope in Eq. (13). We do not run this experiment since planar ground is reasonable assumption for most driving scenarios within some distance.

A1.6. Theorem 1 in Slopy Ground

Theorem 1 remains valid in slopy grounds. The key is the extrapolation behaviour of detectors, and not the ground depth itself.

Proof. Using Eq. (13), the new depth ${}^{g}\hat{z}_{\Delta H}$ is

$$
\begin{aligned}
{}^{g}\hat{z}_{\Delta H} &= \frac{H + \Delta H - b_2 \cos\delta - b_3 \sin\delta}{\frac{v_b + \frac{f\Delta H}{z} - v_0}{f} \cos\delta + \sin\delta} \\
&\approx \frac{H + \Delta H - b_2 \cos\delta - b_3 \sin\delta}{\frac{v_b - v_0}{f} \cos\delta + \sin\delta} \\
&= \frac{H - b_2 \cos\delta - b_3 \sin\delta}{\frac{v_b - v_0}{f} \cos\delta + \sin\delta} + \frac{\Delta H}{\frac{v_b - v_0}{f} \cos\delta + \sin\delta} \\
&= {}^{g}\hat{z}_0 + \frac{\Delta H}{\frac{v_b - v_0}{f} \cos\delta + \sin\delta} \\
&\approx z + \eta + \frac{f\Delta H}{(v_b - v_0)\cos\delta + f \sin\delta} \\
\implies {}^{g}\hat{z}_{\Delta H} - z &\approx \eta + \frac{f\Delta H}{(v_b - v_0)\cos\delta + f \sin\delta},
\end{aligned}
$$

assuming the depth ${}^{g}\hat{z}_0$ at train height $\Delta H = 0$ is the GT depth z added by a normal random variable $\eta(0, \sigma^2)$ [42]. Taking expectation on both sides, mean depth error is

$$
\mathbb{E}\left({}^{g}\hat{z}_{\Delta H} - z\right) \approx \left(\frac{1}{(v_b - v_0)\cos\delta + f \sin\delta}\right) f\Delta H. \quad (14)
$$

Thus, this theorem remains valid for positive slopes and negative slopes $> -\arctan\left(\frac{v_b - v_0}{f}\right)$. The valid slope $\delta \in \left(-\arctan\left(\frac{v_b - v_0}{f}\right), \frac{\pi}{2}\right]$ radians. As an example, for bottom-most point $v_0 = \frac{h}{2}$ and focal length $f = \frac{h}{2}$, the valid delta range $\delta \in \left(-\frac{\pi}{4}, \frac{\pi}{2}\right]$ radians. Since almost all real datasets have slopes $|\delta| < 10$ degrees, the extrapolation behavior remains consistent as Theorem 1. □

A1.7. Unrealistic Assumptions

Flat ground assumption does not hold in real-world datasets. Real datasets like KITTI and nuScenes are mostly flat-ground datasets. Recent methods such as MonoGround [79] and Homography Loss [26] leverage this assumption. CHARM3R makes a crucial **first attempt to address the extrapolation problem** within this relevant and prevalent setting.

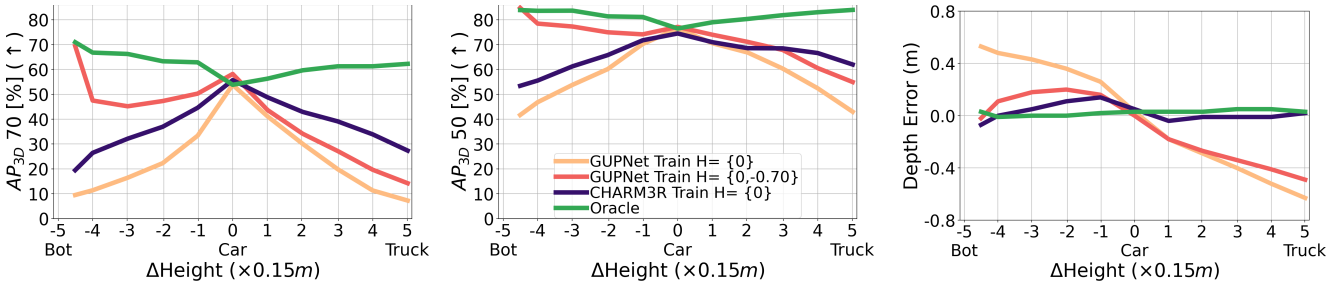
Over-idealized Theorem 2. Car has constant surface depth. Theorem 2 relies on Sec. 3.2 and Fig. 3 of [21], that proves that all depth estimators are regression models. The Mono3D task and previous works: DEVIANT [41] (our baseline) and MonoDETR predict depth at the car center, not its surface. Thus, Theorem 2 addresses regression for this standard center-based depth prediction.

A1.8. Error Trends For Far Objects

Note that the error trends of regression-based depth and ground-based depth do not completely cancel for far objects. However, the baseline Mono3D performance is already notably poor for far objects. Fig. 7 confirms that both regression-based baseline and CHARM3R performance are bad beyond $> 30m$ range.

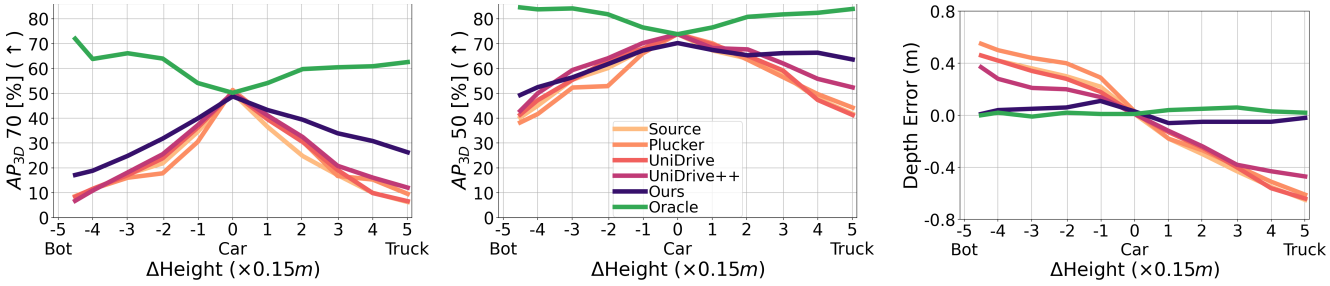
A2. Additional Experiments

We now provide additional details and results of the experiments evaluating CHARM3R's performance.



(a) AP_{3D} 70 [%] comparison. (b) AP_{3D} 50 [%] comparison. (c) MDE comparison.

Figure 8. **CARLA Val Results with GUP Net** detector after augmentation of [38]. Training a detector with both $\Delta H = -0.70m$ and $\Delta H = 0m$ images produces better results at $\Delta H = -0.70m$ and $\Delta H = 0m$, but **fails at unseen height images** $\Delta H = +0.76m$. CHARM3R **outperforms** all baselines, especially at unseen bigger height changes. All methods except Oracle are trained on car height and tested on all heights.



(a) AP_{3D} 70 [%] comparison. (b) AP_{3D} 50 [%] comparison. (c) MDE comparison.

Figure 9. **CARLA Val Results with DEVIANT** detector. CHARM3R **outperforms** all baselines, especially at bigger height changes. All methods except Oracle are trained on car height and tested on all heights. Results of inference on height changes of $-0.70, 0$ and 0.76 meters are in Tab. 2.

Loss Function. CHARM3R uses the same losses as baselines. CHARM3R's final depth estimate, being a fusion of regression and ground-based depth, does not need loss changes.

A2.1. CARLA Val Results

We first analyze the results on the synthetic CARLA dataset further.

AP at different distances and thresholds. We next compare the AP_{3D} of the baseline GUP Net and CHARM3R in Fig. 7 at different distances in meters and IoU_{3D} matching criteria of $0.1 - 0.7$ as in [41]. Fig. 7 shows that CHARM3R is effective over GUP Net at all depths and higher IoU_{3D} thresholds. CHARM3R shows biggest gains on $IoU_{3D} > 0.3$ for $[0, 30]m$ boxes.

Comparison with Augmentation-Methods. Sec. 1 of the paper says that the augmentation strategy falls short when the target height is OOD. We show this in Fig. 8. Since authors of [38] do not release the NVS code, we use the ground truth images from height change $\Delta H = -0.70m$ in training. Fig. 8 confirms that augmentation also improves the performance on $\Delta H = -0.70m$ and $\Delta H = 0m$, but again falls short on unseen ego heights $\Delta H = +0.76m$. On the other hand, CHARM3R (even though trained on $\Delta H = -0.70m$) outperforms such augmentation strategy at unseen ego heights $\Delta H = +0.76m$. This shows the complementary nature of CHARM3R over augmentation strategies.

Reproducibility. We ensure reproducibility of our results by repeating our experiments for 3 random seeds. We choose the final epoch as our checkpoint in all our experiments as [41, 42]. Tab. 5 shows the results with these seeds. CHARM3R outperforms the baseline GUP Net in both median and average cases.

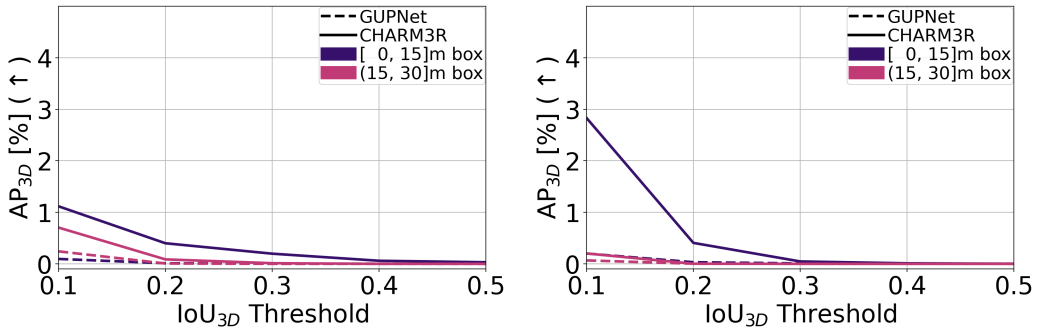
Results with DEVIANT. We next additionally plot the robustness of CHARM3R with other methods on the DEVIANT detector [41] in Fig. 9 The figure confirms that CHARM3R works even with DEVIANT and produces SoTA robustness to unseen ego heights.

A2.2. nuScenes ➙ CODa Val Results

To test our claims further in real-life, we use two real datasets: the nuScenes dataset [7] and the recently released CODa [115] datasets. nuScenes has ego camera at height $1.51m$ above the ground, while the CODa is a robotics dataset with ego camera at a height of $0.75m$ above the

3D Detector	Seed ↓ / ΔH (m)→	AP_{3D} 70 [%] (↑)			AP_{3D} 50 [%] (↑)			MDE (m) [≈ 0]		
		-0.70	0	$+0.76$	-0.70	0	$+0.76$	-0.70	0	$+0.76$
GUP Net [62]	111	12.24	55.98	7.53	44.14	76.37	41.32	+0.48	+0.00	−0.64
	444	9.46	53.82	7.23	41.66	76.47	40.97	+0.53	+0.03	−0.63
	222	10.35	52.94	10.79	41.67	75.80	46.45	+0.53	+0.01	−0.57
	Average	10.68	54.25	8.52	42.49	**76.21**	43.58	+0.51	+0.01	−0.61
+ CHARM3R	111	19.99	58.16	29.96	54.15	74.10	64.27	+0.09	+0.00	−0.03
	444	19.45	55.68	27.33	53.40	74.47	61.98	+0.07	+0.05	−0.02
	222	17.41	53.57	27.77	54.30	74.83	64.42	+0.12	+0.01	−0.09
	Average	**18.95**	**55.80**	**28.35**	**53.95**	74.47	**63.56**	+0.09	+0.02	−0.05
Oracle	–	70.96	53.82	62.25	83.88	76.47	83.96	+0.03	+0.03	+0.03

Table 5. **Reproducibility Results.** CHARM3R **outperforms** all other baselines on CARLA Val split, especially at bigger unseen ego heights in both median (Seed=444) and average cases. All except Oracle are trained on car height $\Delta H = 0m$ and tested on bot to truck height data. [Key: **Best**]



(a) CODa Car (b) CODa Pedestrian

Figure 10. **CODa Val AP_{3D} at different depths and IoU_{3D} thresholds** with GUP Net trained on nuScenes. CHARM3R shows biggest gains on $IoU_{3D} < 0.3$ for $[0, 30]m$ boxes.

Val	**Ego Ht** (m)	**#Images**	**Car** (k)	**Ped** (k)
nuScenes	1.51	6,019	18	7
CODa	0.75	4,176	4	86

Table 6. **Dataset statistics.** nuScenes Val has more Cars compared to Pedestrians, while CODa Val has more Pedestrians than Cars.

ground. Tab. 6 shows the statistics of these two datasets. This experiment uses the following data split:

- *nuScenes Val Split.* This split [7] contains 28,130 training and 6,019 validation images from the front camera as [41].
- *CODa Val Split.* This split [115] contains 19,511 training and 4,176 validation images. We only use this split for testing.

We train the GUP Net detector with 10 nuScenes classes and report the results with the KITTI metrics on both nuScenes val and CODa Val splits.

Main Results. We report the main results in Tab. 7 paper. The results of Tab. 7 shows gains on both Cars and Pedestrians classes of CODa val dataset. The performance is very low, which we believe is because of the domain gap between nuScenes and CODa datasets. These results further confirm our observations that unlike 2D detection, generalization across unseen datasets remains a big problem in the Mono3D task.

AP at different distances and thresholds. To further analyze the performance, we next plot the AP_{3D} of the baseline GUP Net and CHARM3R in Fig. 10 at different distances in meters and IoU_{3D} matching criteria of $0.1 - 0.5$ as in [41]. Fig. 10 shows that CHARM3R is effective over GUP Net at all depths and lower IoU_{3D} thresholds. CHARM3R shows biggest gains on $IoU_{3D} < 0.3$ for $[0, 30]m$ boxes. The gains are more on the Pedestrian class on CODa since CODa captures UT Austin campus scenes, and therefore, has more pedestrians compared to cars. nuScenes captures outdoor driving scenes in Boston and Singapore, and therefore, has more cars compared to pedestrians. We describe the statistics of these two datasets in Tab. 6.

A2.3. Qualitative Results.

CARLA. We now show some qualitative results of models trained on CARLA Val split from car height $(\Delta H = 0m)$ and tested on truck height $(\Delta H = +0.76m)$ in Fig. 11. We depict the predictions of CHARM3R in image view on the left, the predictions of CHARM3R, the baseline GUP

3D Detector	Method	Car AP_{3D} 50 [%] (↑)		Ped AP_{3D} 30 [%] (↑)	
		CODa	nuScenes	CODa	nuScenes
GUP Net [62]	Source	0.02	**18.42**	0.01	**2.93**
	UniDrive [47]	0.02	**18.42**	0.01	**2.93**
	UniDrive++[47]	**0.03**	**18.42**	**0.02**	**2.93**
	CHARM3R	**0.30**	**14.80**	**0.05**	**1.26**
	Oracle	28.56	18.42	30.31	2.93

Table 7. **nuScenes to CODa Val Results.** CHARM3R **outperforms** all baselines, especially at unseen height changes. [Key: **Best**, **Second Best**, Ped= Pedestrians]

Net [62], and GT boxes in BEV on the right. In general, CHARM3R detects objects more accurately than GUP Net [62], making CHARM3R more robust to camera height changes. The regression-based baseline GUP Net mostly underestimates the depth of 3D boxes with positive ego height changes, which qualitatively justifies the claims of Theorem 2.

CODa. We now show some qualitative results of models trained on CODa Val split in Fig. 12. As before, we depict the predictions of CHARM3R in image view image view on the left, the predictions of CHARM3R, the baseline GUP Net [62], and GT boxes in BEV on the right. In general, CHARM3R detects objects more accurately than the baseline GUP Net [62], making CHARM3R more robust to camera height changes. Also, considerably less number of boxes are detected in the cross-dataset evaluation *i.e.* on CODa Val. We believe this happens because of the domain shift.

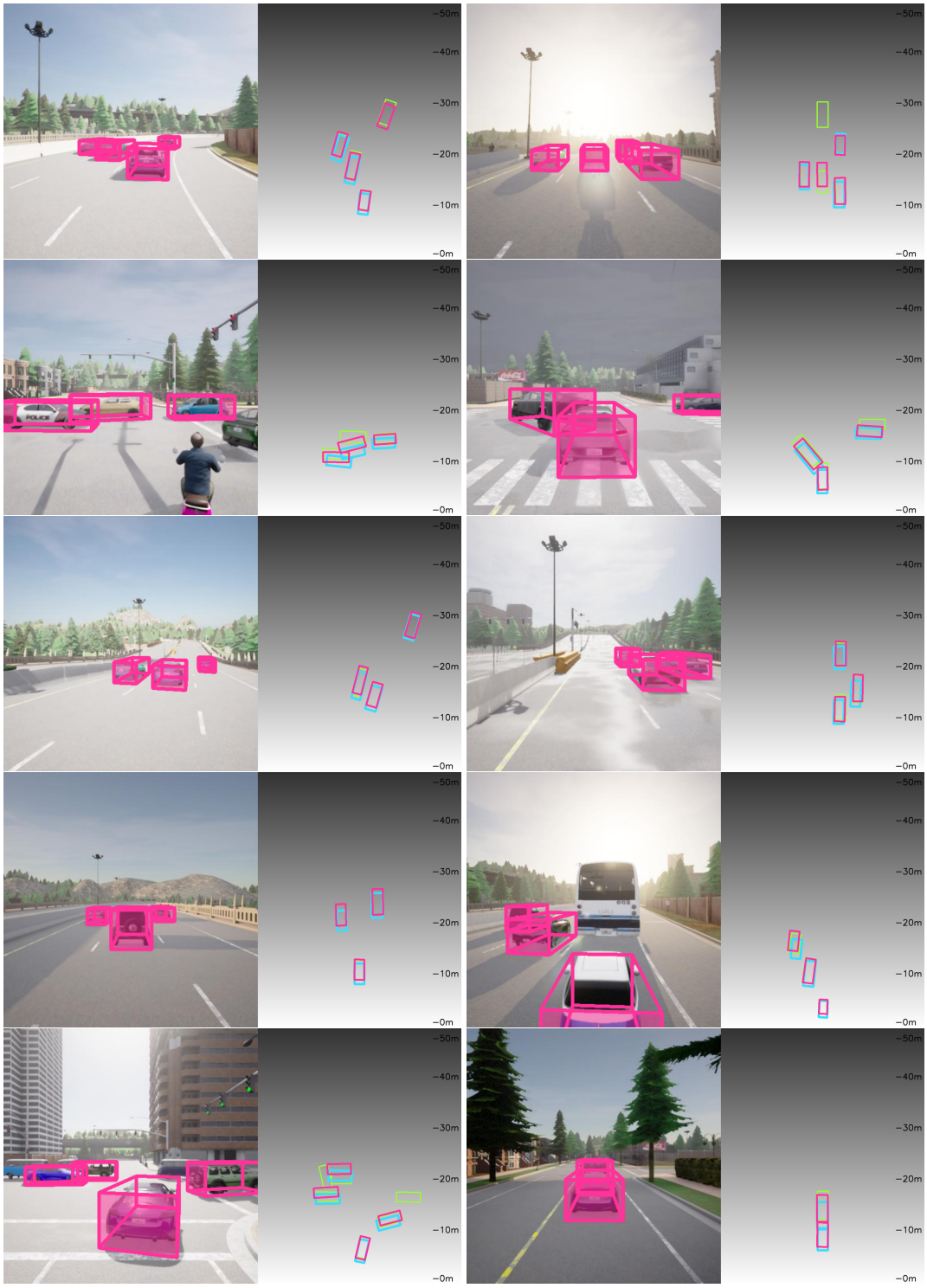


Figure 11. **CARLA Val Qualitative Results**. CHARM3R detects objects more accurately than GUP Net [62], making CHARM3R more robust to camera height changes. The regression-based baseline GUP Net mostly underestimates the depth which qualitatively justifies the claims of Theorem 2. All methods are trained on CARLA images at car height $\Delta H = 0m$ and evaluated on $\Delta H = +0.76m$. [Key: Cars of CHARM3R. ; Cars of GUP Net, and Ground Truth in BEV.

Figure 12. **CODa Val Qualitative Results**. CHARM3R detects objects more accurately than GUP Net [62], making CHARM3R more robust to camera height changes. All methods are trained on nuScenes dataset and evaluated on CODa dataset. [Key: Cars and Pedestrian of CHARM3R. ; all classes of GUP Net, and Ground Truth in BEV.