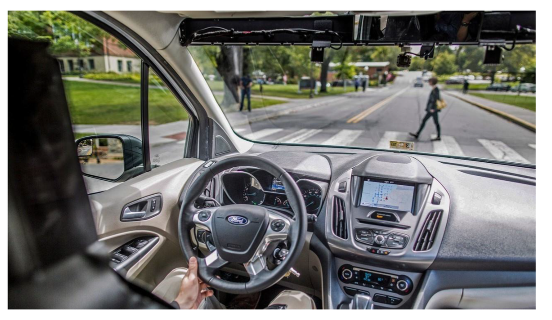
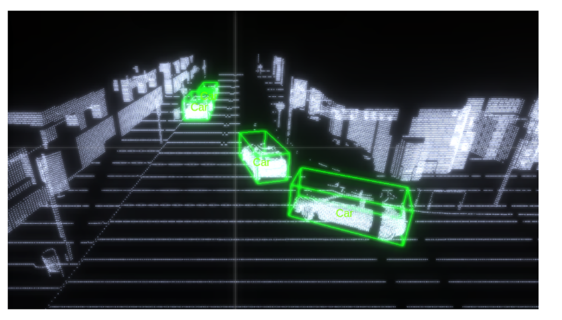
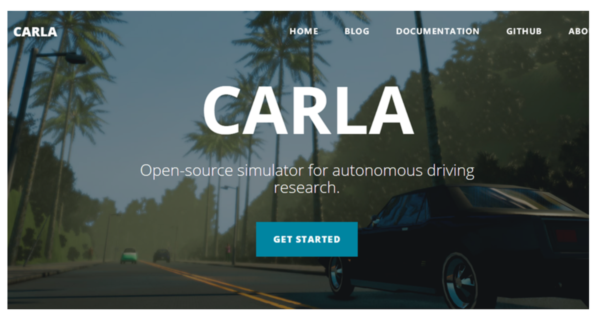
Manual-Label Free 3D Detection via An Open-Source Simulator

Zhen Yang, Chi Zhang, Huiming Guo, and Zhaoxiang Zhang
Institute of Automation, Chinese Academy of Sciences (CASIA)
Beijing Aerospace Changfeng Co.Ltd., The 2nd Institute of CASIC

Background

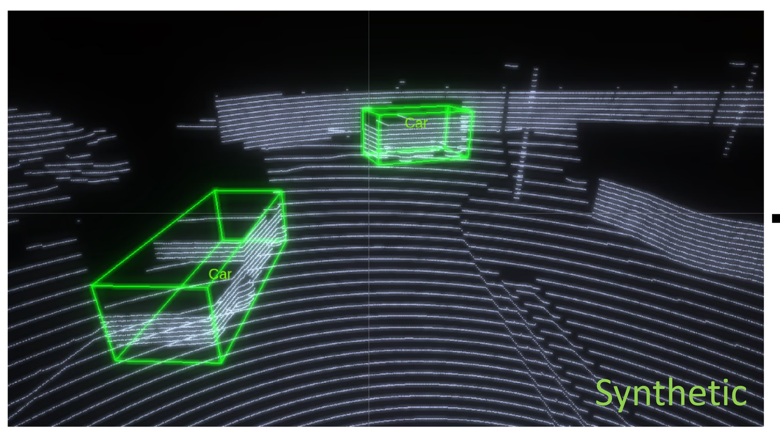




Manual labeled point cloud data is scarce and expensive.

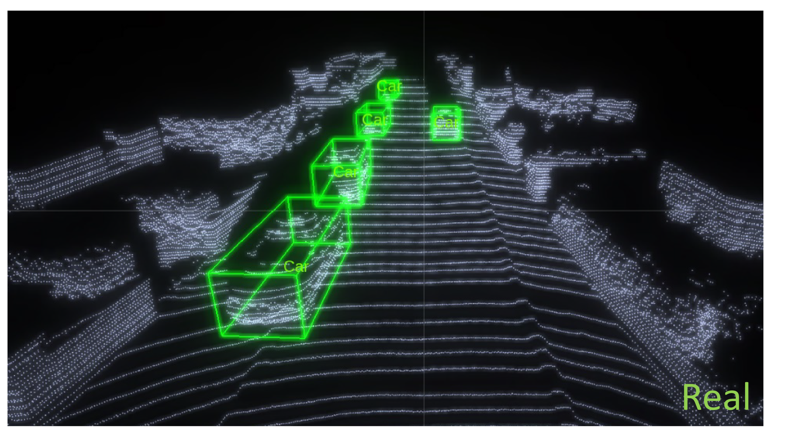



Recently, the simulators are being increasingly used to remedy the shortage of labeled data.

Challenges



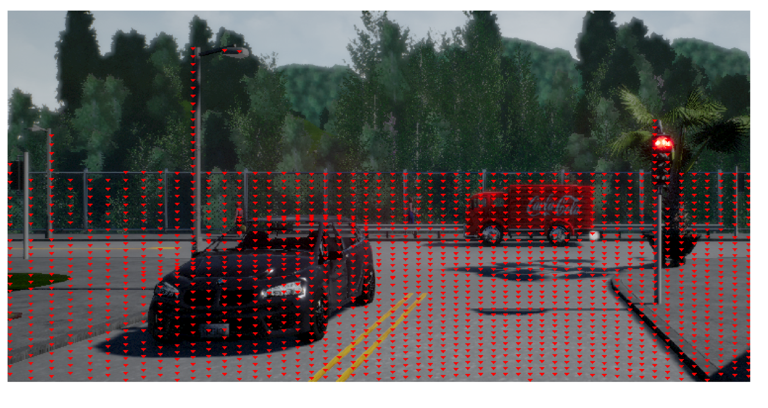
Gap



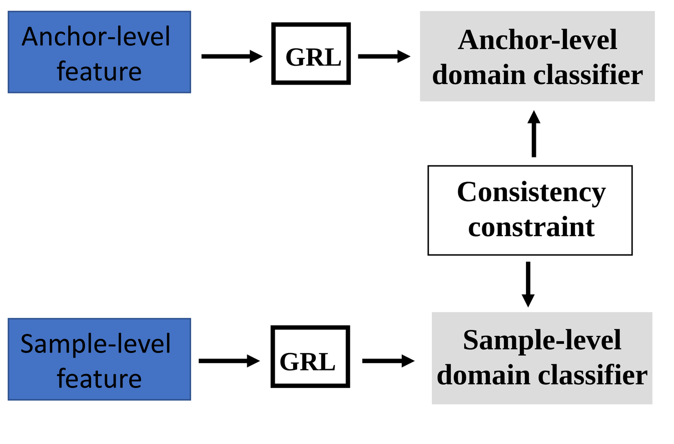
The synthetic data is severely distorted, and such discrepancies would cause significant performance drop.

Our Methods

LiDAR-guided Sampling: Virtual lidar helps reduce the distortion of the synthetic point cloud data.



Domain adaptation: Feature alignment helps reduce the impact of the discrepancy between the synthetic data and the real data on model performance.

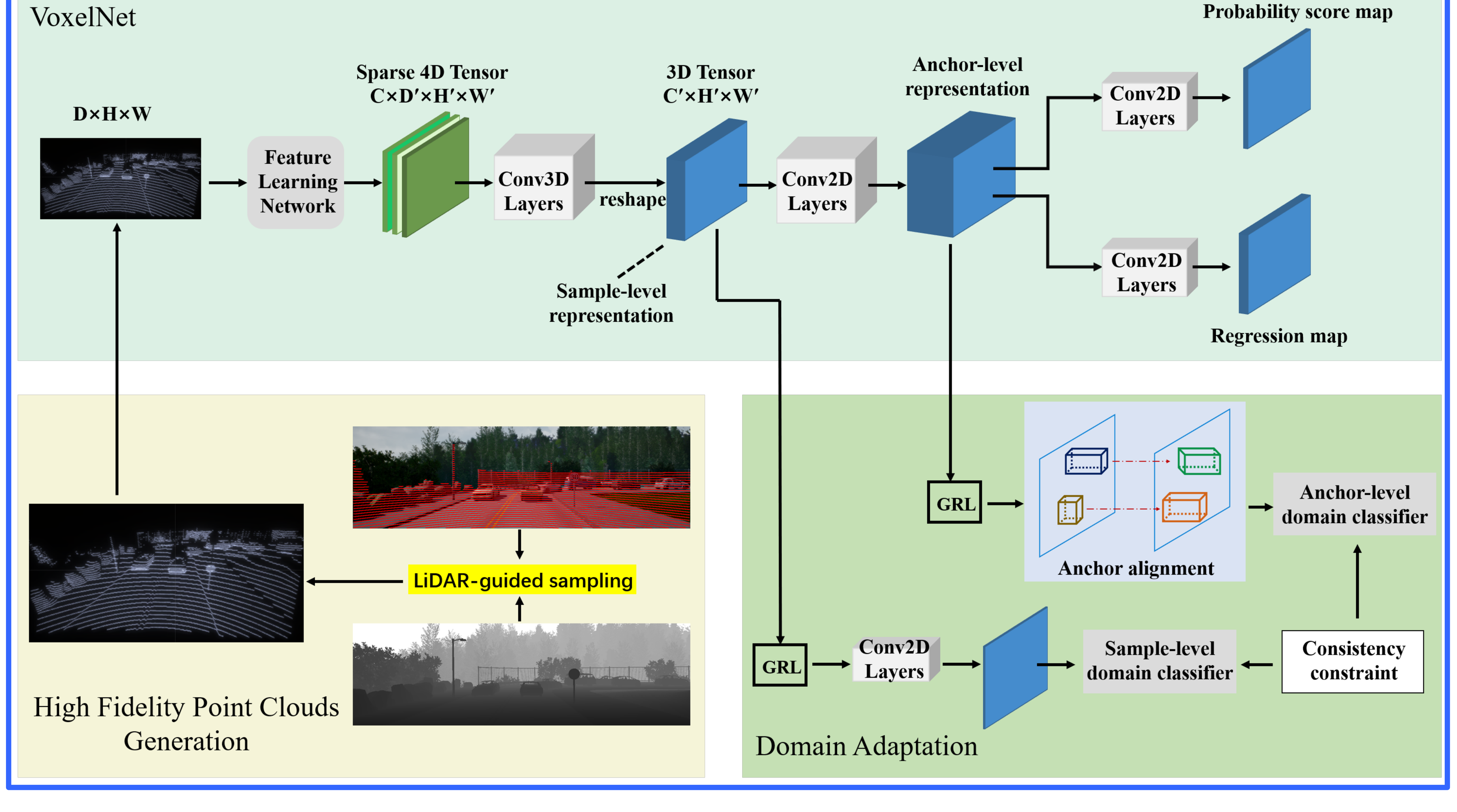


Contribution

1) We produce high quality 3D models and embed these models into CARLA simulator to get more realistic virtual point clouds. We then propose a novel sampling algorithm, LiDAR-guided sampling, to generate high fidelity point cloud samples. By utilizing the high fidelity point clouds to augment the training set, we can implement a promising 3D detector with exponentially reduced manual labeled data.

2) We propose two novel domain adaptation components to cross the gap between the synthetic data and the real data. We further impose a consistency constraint to stabilize the training process. Combine the both themes and based on the 3D detector VoxelNet, the proposed DA-VoxelNet can get rid of the manual annotations thoroughly.

Framework



Algorithm and Objective Function

LiDAR-guided sampling:

$$f(x,y,D)=\begin{cases}1, & (x,y)\in D\\ 0, & (x,y)\notin D\end{cases}$$

$$\delta(d)=\begin{cases}1, & d\geq d_{min}\\ 0, & d<d_{min}\end{cases}$$

$$f(u,v,P_s)=\iint_{HW} f(x,y,D_{aug})\delta(\|(x,y),(u,v)\|_2)dxdy$$

$$P=D_s\cup P_s$$

Sample-Level Adaptation:

$$L_{s_s}=-\frac{1}{n_s}\sum_{i=1}^{n_s}\log(1-D_s(F_s(p_i^s)))$$

$$L_{s_t}=-\frac{1}{n_t}\sum_{i=1}^{n_t}\log(D_s(F_s(p_i^t)))$$

$$L_{sample}=L_{s_s}+L_{s_t}$$

Anchor-Level Adaptation:

$$L_{ac_s}=-\frac{1}{n_s}\sum_{i=1}^{n_s}\log(1-D_a(F_a(p_i^s)))$$

$$L_{ac_t}=-\frac{1}{n_t}\sum_{i=1}^{n_t}\log(D_a(F_a(p_i^t)))$$

$$L_{anchor}=L_{ac_s}+L_{ac_t}$$

Consistency Constraint:

$$M_s(n,p_i)=\frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H_s}\sum_{w=1}^{W_s}D_s(F_s(p_i))_{(w,h)}$$

$$M_a(n,p_i)=\frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H_a}\sum_{w=1}^{W_a}D_a(F_a(p_i))_{(w,h)}$$

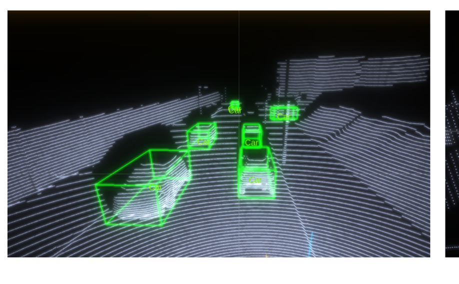
$$L_{con_f}(n,p_i)=\|M_s(n,p_i)-M_a(n,p_i)\|_2$$

$$L_{con}=L_{con_f}(n_s,p_i^s)+L_{con_f}(n_t,p_i^t)$$

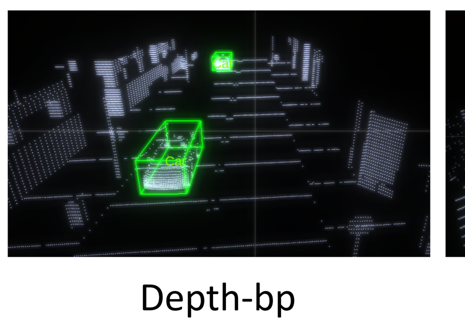
Overall Objective:

$$L=L_{det}+\lambda(L_{sample}+L_{anchor}+L_{con})$$

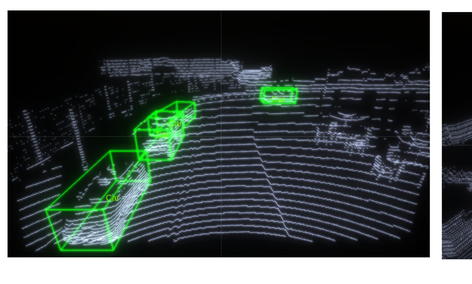
Experimental Results



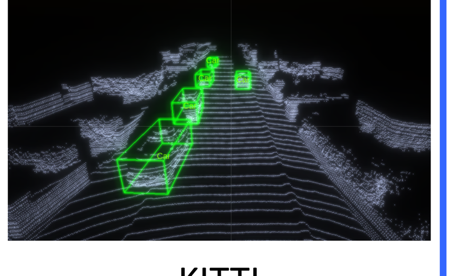
CARLA-origin



Depth-bp



LiDAR-guided



KITTI

TABLE I: The average precision (AP) of Car on the KITTI validation set and nuScenes validation set respectively. The VoxelNet is trained using the training set of LIDAR dataset (L), DEPTH dataset (D), CARLA dataset (C), KITTI (K) and nuScenes (nuS) as the source domain respectively. Among them, L, D and C are synthetic dataset generated by the CARLA simulator, K and nuS are collected from the real scene. **Red** indicates the best and **Blue** the second best.

	Direction	Easy	Moderate	Hard	Direction	Easy	Moderate	Hard
BEV AP	C→K	61.09	53.13	49.30	C→nuS	37.11	31.27	14.41
	D→K	83.71	71.56	66.22	D→nuS	50.82	40.75	19.40
	L→K	**87.71**	**74.89**	**68.01**	L→nuS	**55.46**	**46.30**	**20.21**
	K→K	**89.97**	**87.85**	**86.84**	nuS→nuS	**74.82**	**65.99**	**31.67**
3D AP	C→K	26.64	21.98	20.56	C→nuS	2.08	1.79	1.30
	D→K	68.09	52.84	46.00	D→nuS	**25.51**	**20.12**	**10.53**
	L→K	**71.78**	**55.03**	**47.42**	L→nuS	23.78	18.32	9.75
	K→K	**88.41**	**78.37**	**77.33**	nuS→nuS	**49.82**	**42.24**	**20.54**

TABLE II: Quantitative analysis of finetune result from synthetic data to real data. *Percentage* denotes the number of sampled data as a percentage of the target training set. If *Finetune*, we use the synthetic data to train the model and use the sampled data to finetune the model, else we use the sampled data to train the model directly.

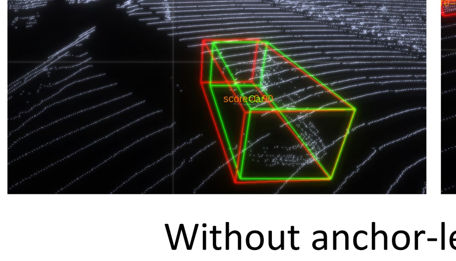
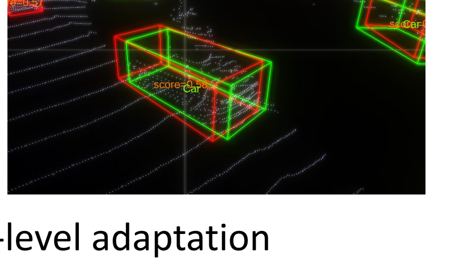
	Percentage	Finetune	Direction	Easy	Moderate	Hard	Direction	Easy	Moderate	Hard
BEV AP	1%	×	K→K	29.12	23.61	18.08	nuS→nuS	34.52	27.54	13.96
	1%	✓	L→K	89.49	79.22	77.95	L→nuS	70.88	61.70	29.25
	5%	✓	L→K	89.75	85.68	79.33	L→nuS	72.80	63.78	29.79
	10%	✓	L→K	**90.24**	86.71	86.31	L→nuS	73.94	64.67	30.19
	100%	×	K→K	89.97	**87.85**	**86.84**	nuS→nuS	**74.82**	**65.99**	**31.67**
3D AP	1%	×	K→K	10.72	10.65	7.57	nuS→nuS	7.21	5.03	2.49
	1%	✓	L→K	83.82	72.25	66.00	L→nuS	39.04	29.98	15.35
	5%	✓	L→K	87.02	75.55	68.45	L→nuS	46.26	38.15	18.53
	10%	✓	L→K	87.51	76.40	74.45	L→nuS	49.44	41.27	19.60
	100%	×	K→K	**88.41**	**78.37**	**77.33**	nuS→nuS	**49.82**	**42.24**	**20.54**

TABLE III: Results on adaptation from LIDAR to KITTI Dataset. Average precision (AP) of Car is evaluated on the KITTI validation set. *bs* is short for batch size.

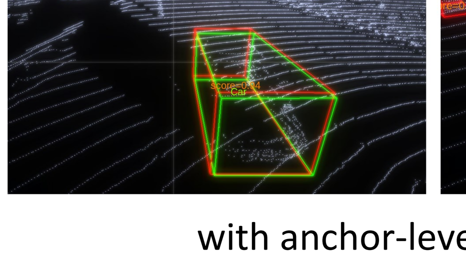
	bs	method	Easy	Moderate	Hard
BEV AP	2	VoxelNet	79.27	66.72	63.33
		DA-VoxelNet	81.19	71.27	65.18
	8	VoxelNet	87.71	74.89	68.01
		DA-VoxelNet	**88.40**	**76.66**	**74.07**
3D AP	2	VoxelNet	57.04	43.02	40.62
		DA-VoxelNet	65.18	51.61	45.10
	8	VoxelNet	71.78	55.03	47.42
		DA-VoxelNet	**73.77**	**56.64**	**52.29**

TABLE IV: Ablation study: Quantitative results on the KITTI validation set for *Moderate* level, reported as mean and standard deviation over 3 rounds of training with batch size 2. Models are trained on the LIDAR training set. *an* is short for anchor-level adaptation, *sa* for sample-level adaptation and *cons* is short for our consistency constraint.

method	sa	an	cons	BEV AP (mean±std)	3D AP (mean±std)
VoxelNet				66.44±0.43	43.42±0.62
DA-VoxelNet	✓			69.49±1.70	48.12±1.16
		✓		70.50±0.20	50.30±0.42
	✓	✓		70.92±0.17	50.15±0.68
	✓	✓	✓	**71.15±0.33**	**50.57±1.61**

Without anchor-level adaptation




with anchor-level adaptation

Conclusions

- LiDAR-guided sampling is helpful.
  - ➢ The high fidelity point cloud samples obtained by using LiDAR-guided sampling method can improve the detector's generalization ability on real scenes.
- DA-VoxelNet can relieve the domain gap.
  - ➢ DA-VoxelNet gain a large performance improvement compared to the VoxelNet, which reveals a promising perspective of training a LiDAR-based 3D detector without any hand-tagged label.