

VirConv Appendix

1 Code and Data Licence

We released our code under “Apache License 2.0”. The KITTI Dataset and nuScenes Dataset are both licensed for academic research.

2 More Results on KITTI leaderboard

3D detection performance. We show a screenshot of KITTI’s 3D detection online leaderboard in Fig. 1, where our results are highlighted in red boxes. Our VirConv-S, VirConv-T and VirConv-L achieve high detection performance and currently rank 1st, 2nd and 5th, respectively. The outstanding performance comes from that our new StVD and NRConv designs better leverage the geometry clue from virtual points, leading to high-quality 3D object detection.

Car


	Method	Setting	Code	Moderate	Easy	Hard	Runtime
1	VirConv-S			87.20 %	92.48 %	82.45 %	0.09 s
2	VirConv-T			86.25 %	92.54 %	81.24 %	0.09 s
3	TED			85.28 %	91.61 %	80.68 %	0.1 s
4	LoGoNet			85.06 %	91.80 %	80.74 %	0.1 s
5	VirConv-L			85.05 %	91.41 %	80.22 %	0.05 s
6	LIVOX_Det			84.94 %	91.72 %	80.10 %	n/a s
7	SFD		code	84.76 %	91.73 %	77.92 %	0.1 s
X. Wu, L. Peng, H. Yang, L. Xie, C. Huang, C. Deng, H. Liu and D. Cai: Sparse Fuse Dense: Towards High Qu							
8	VoCo			84.76 %	91.99 %	79.81 %	0.1 s
9	NSAW		code	84.30 %	90.57 %	77.46 %	0.1 s
10	CasA++		code	84.04 %	90.68 %	79.69 %	0.1 s
H. Wu, J. Deng, C. Wen, X. Li and C. Wang: CasA: A Cascade Attention Network for 3D Object Detection frc Remote Sensing 2022.							
11	DGDNH			83.88 %	90.69 %	79.50 %	0.04 s
12	Anonymous			83.51 %	89.08 %	78.94 %	n/a s
13	GraR-Vol		code	83.27 %	91.89 %	77.78 %	0.07 s
H. Yang, Z. Liu, X. Wu, W. Wang, W. Qian, X. He and D. Cai: Graph R-CNN: Towards Accurate 3D Object Det							
14	GLENet-VR			83.23 %	91.67 %	78.43 %	0.04 s
Y. Zhang, Q. Zhang, Z. Zhu, J. Hou and Y. Yuan: GLENet: Boosting 3D Object Detectors with Generative Lab							
15	VPFNet		code	83.21 %	91.02 %	78.20 %	0.06 s
H. Zhu, J. Deng, Y. Zhang, J. Ji, Q. Mao, H. Li and Y. Zhang: VPFNet: Improving 3D Object Detection with V Transactions on Multimedia 2022.							

Figure 1: KITTI 3D Car detection leaderboard, where we only show the top 15 results among 448 submissions. Our VirConv-S, VirConv-T and VirConv-L rank 1st, 2nd and 5th among all submissions, respectively.

BEV detection performance. Besides, our methods achieved the highest AP in the BEV detection leaderboard. We show the results in Fig. 2. Our VirConv-S, VirConv-T and VirConv-L rank 1st, 2nd and 6th, respectively. The reason is that our method can boost the object localization accuracy and reduce the noise impact of virtual points.

Car



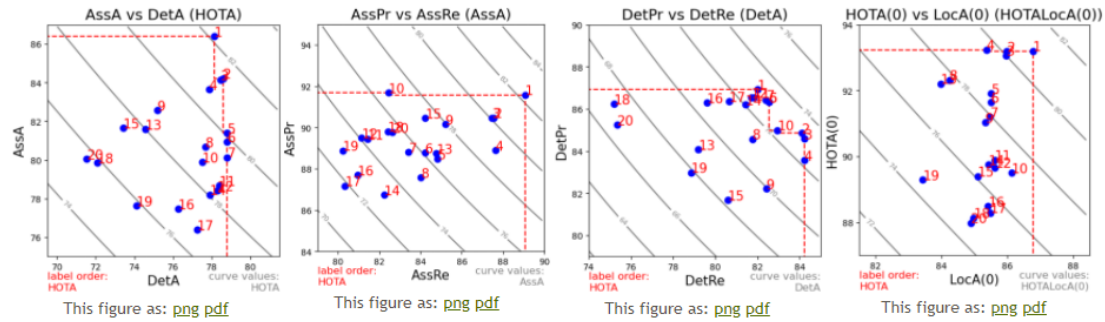
	Method	Setting	Code	Moderate	Easy	Hard	Runtime
1	VirConv-S			93.52 %	95.99 %	90.38 %	0.09 s
2	VirConv-T			92.65 %	96.11 %	89.69 %	0.09 s
3	GraR-Po		code	92.12 %	95.79 %	87.11 %	0.06 s
	H. Yang, Z. Liu, X. Wu, W. Wang, W. Qian, X. He and D. Cai: Graph R-CNN: Towards Accurate 3D Object Det						
4	TED			92.05 %	95.44 %	87.30 %	0.1 s
5	LIVOX Det			92.05 %	95.60 %	89.22 %	n/a s
6	VirConv-L			91.95 %	95.53 %	87.07 %	0.05 s
7	VPFNet		code	91.86 %	93.02 %	86.94 %	0.06 s
	H. Zhu, J. Deng, Y. Zhang, J. Ji, Q. Mao, H. Li and Y. Zhang: VPFNet: Improving 3D Object Detection with V Transactions on Multimedia 2022.						
8	SFD		code	91.85 %	95.64 %	86.83 %	0.1 s
	X. Wu, L. Peng, H. Yang, L. Xie, C. Huang, C. Deng, H. Liu and D. Cai: Sparse Fuse Dense: Towards High Que						
9	SE-SSD		code	91.84 %	95.68 %	86.72 %	0.03 s
	W. Zheng, W. Tang, L. Jiang and C. Fu: SE-SSD: Self-Ensembling Single-Stage Object Detector From Point Cl						
10	GraR-Vo		code	91.72 %	95.27 %	86.51 %	0.04 s
	H. Yang, Z. Liu, X. Wu, W. Wang, W. Qian, X. He and D. Cai: Graph R-CNN: Towards Accurate 3D Object Det						
11	PVT-SSD			91.63 %	95.23 %	86.43 %	0.05 s
12	CityBrainLab			91.62 %	94.78 %	86.68 %	0.04 s
13	SPANet			91.59 %	95.59 %	86.53 %	0.06 s
	Y. Ye: SPANet: Spatial and Part-Aware Aggregation Network for 3D Object Detection . Pacific Rim Internatio						
14	CasA		code	91.54 %	95.19 %	86.82 %	0.1 s
	H. Wu, J. Deng, C. Wen, X. Li and C. Wang: CasA: A Cascade Attention Network for 3D Object Detection fro Remote Sensing 2022.						
15	LoGoNet			91.52 %	95.48 %	87.09 %	0.1 s

Figure 2: KITTI BEV detection leaderboard, where we only show the top 15 results among 466 submissions. Our VirConv-S, VirConv-T and VirConv-L rank 1st, 2nd and 6th among all submissions, respectively.

Multi-object tracking performance. In addition, our methods can be easily extended to other downstream tasks, such as object tracking. To demonstrate it, we constructed a simple tracking-by-detection framework, named VirConvTrack. We first detect all 3D objects from point clouds using our VirConv-T. Then we associate the objects between frames based on the Kalman filtering as similar as [3]. We evaluate our tracking performance on the KITTI tracking benchmark. The results are shown in Fig. 3. Our VirConvTrack ranks 1st among all past submissions on the leaderboard, demonstrating the great generalization ability of our method.

CAR



Method	Setting	Code	HOTA	DetA	AssA	DetRe	DetPr	AssRe	AssPr	LocA	MOTA	Compare
1	VirConvTrack		81.87 %	78.14 %	86.39 %	82.00 %	86.92 %	89.08 %	91.58 %	88.04 %	90.24 %	<input type="checkbox"/>
2	CasTrack		81.00 %	78.58 %	84.22 %	84.10 %	84.86 %	87.55 %	90.47 %	87.49 %	91.91 %	<input type="checkbox"/>
H. Wu, J. Deng, C. Wen, X. Li and C. Wang: CasA: A Cascade Attention Network for 3D Object Detection from LiDAR point clouds . IEEE TGRS 2022.												
H. Wu, W. Han, C. Wen, X. Li and C. Wang: 3D Multi-Object Tracking in Point Clouds Based on Prediction Confidence-Guided Data Association . IEEE TITS 2021.												
3	PC-TCNN		80.90 %	78.46 %	84.13 %	84.22 %	84.58 %	87.46 %	90.47 %	87.48 %	91.70 %	<input type="checkbox"/>
H. Wu, Q. Li, C. Wen, X. Li, X. Fan and C. Wang: Tracklet Proposal Network for Multi-Object Tracking on Point Clouds . IJCAI 2021.												
4	Rethink MOT		80.39 %	77.88 %	83.64 %	84.23 %	83.57 %	87.63 %	88.90 %	87.07 %	91.53 %	<input type="checkbox"/>
5	CMOT-RAM-DeepSort		79.76 %	78.80 %	81.40 %	82.55 %	86.33 %	84.81 %	88.47 %	87.15 %	91.62 %	<input type="checkbox"/>
6	RAM		79.53 %	78.79 %	80.94 %	82.54 %	86.33 %	84.21 %	88.77 %	87.15 %	91.61 %	<input type="checkbox"/>
P. Tokmakov, A. Jabri, J. Li and A. Gaidon: Object Permanence Emerges in a Random Walk along Memory . ICML 2022.												
7	Anonymous		79.13 %	78.81 %	80.13 %	82.41 %	86.43 %	83.40 %	88.81 %	87.11 %	91.72 %	<input type="checkbox"/>
8	FastTrack		78.78 %	77.67 %	80.66 %	81.76 %	84.57 %	84.02 %	87.58 %	86.01 %	92.06 %	<input type="checkbox"/>
9	MSA-MOT		78.52 %	75.19 %	82.56 %	82.42 %	82.21 %	85.21 %	90.16 %	87.00 %	88.01 %	<input type="checkbox"/>
10	CyberTrack		78.25 %	77.51 %	79.88 %	82.95 %	84.99 %	82.45 %	91.69 %	87.62 %	90.14 %	<input type="checkbox"/>
11	CMOT Perma-Perma		78.21 %	78.39 %	78.67 %	81.81 %	86.53 %	81.43 %	89.44 %	87.10 %	91.50 %	<input type="checkbox"/>
12	PermaTrack		78.03 %	78.29 %	78.41 %	81.71 %	86.54 %	81.14 %	89.49 %	87.10 %	91.33 %	<input type="checkbox"/>
P. Tokmakov, J. Li, W. Burgard and A. Gaidon: Learning to Track with Object Permanence . ICCV 2021.												
13	PC3T		77.80 %	74.57 %	81.59 %	79.19 %	84.07 %	84.77 %	88.75 %	86.07 %	88.81 %	<input type="checkbox"/>
H. Wu, W. Han, C. Wen, X. Li and C. Wang: 3D Multi-Object Tracking in Point Clouds Based on Prediction Confidence-Guided Data Association . IEEE TITS 2021.												
14	StrongSORT++		77.75 %	77.89 %	78.20 %	81.42 %	86.22 %	82.24 %	86.73 %	86.96 %	90.35 %	<input type="checkbox"/>
15	jerrymot		77.12 %	73.43 %	81.66 %	80.60 %	81.69 %	84.23 %	90.45 %	86.79 %	85.82 %	<input type="checkbox"/>

Figure 3: KITTI Car tracking leaderboard, where we only show the top 15 results among 99 submissions. Our VirConvTrack ranks 1st among all submissions.

3 More Experiments

NRConv: simple concatenation vs. weighted fusion. In our main paper section 3.3, we directly concatenate the 3D geometry features and 2D noise-aware features for 3D detection. Here we provide an alternative method of weighted fusion. The comparison results are shown in Table 1. We observe that the simple concatenation works slightly better than the weighted fusion. Thus we adopted the concatenation in our paper.

Table 1: Ablation study results using different fusion methods in NRConv.

fusion method	3D AP		
	Easy	Mod.	Hard
weighted fusion	93.06	88.45	85.55
simple concatenation	95.36	88.71	85.83

Table 2: VirConv-S 3D detection results using different training schemes (on KITTI validation set).

Training scheme for VirConv-S	3D AP		
	Easy	Moderate	Hard
Train from scratch (65 epoch)	95.49	90.44	90.73
Fine tuning (5 epoch)	95.66	90.42	89.66
Fine tuning (10 epoch)	95.76	90.97	89.14
Fine tuning (20 epoch)	95.46	90.37	90.68

VirConv-S: fine-tuning vs. train from scratch. Our semi-supervised VirConv-S can be trained by fine-tuning a pre-trained model or trained from scratch. We conducted several experiments to examine the performance of different training schemes and choose the best scheme. The results are shown in Table 2. We observe that the fine-tuning based training scheme obtains the best results. The reason may be that the training from scratch easily overfits the noisy pseudo labels. The pre-trained model in the fine-tuning scheme is trained more epochs on real labels, preventing the network from overfitting. Thus, we adopted the fine-tuning scheme to train our VirConv-S.

Evaluation with single-stage detection. Generally, the single-stage detector runs much faster than the two-stage detectors. We also conducted an experiment to test our VirConv-L with a single-stage. We reported the results in Table 3. We observe that, with our VirConv, the detection accuracy and efficiency are significantly improved by 5.41% AP and 42 ms, respectively. The results demonstrate that our VirConv can also be generalized to single-stage 3D detectors.

Table 3: Evaluation with the single-stage setting. The RH denotes the refinement head.

VirConv-T w/o RH	3D AP			Time (ms)
	Easy	Mod.	Hard	
w/o VirConv	91.23	81.03	79.12	80
w VirConv	92.71	86.44	83.06	38

Table 4: Ablation study results using different sampling strategies in StVD.

Sampling method	3D AP		
	Easy	Mod.	Hard
Random sampling	92.44	85.67	83.90
Bin-based sampling	95.36	88.71	85.83

Table 5: The multi-class results on the nuScenes test set. ‘C.V.’, ‘Ped.’, and ‘T.C.’ are short for construction vehicle, pedestrian, and traffic cone, respectively. ‘L’ and ‘C’ represent LiDAR and Camera, respectively. VP denotes virtual points.

Method	Modality	mAP	NDS	Car	Truck	C.V.	Bus	Trailer	Barrier	Motor.	Bike	Ped.	T.C.
CenterPoint + VP	LC	66.4	70.5	86.8	58.5	26.1	67.4	57.3	74.8	70.0	49.3	89.1	85.0
CenterPoint + VP + VirConv	LC	67.2	71.2	87.6	59.7	28.8	68.0	58.2	75.1	70.3	49.7	89.2	85.3
TransFusion	LC	68.9	71.7	87.1	60.0	33.1	68.3	60.8	78.1	73.6	52.9	88.4	86.7
TransFusion-L+VP	LC	66.7	70.8	87.2	58.2	28.8	67.9	61.7	74.8	69.3	45.8	88.2	85.0
TransFusion-L +VP + VirConv	LC	68.7	72.3	88.1	60.3	31.0	69.9	63.3	75.6	75.0	50.3	88.3	85.5

StVD: sampling strategy comparison. In our main paper section 3.2, we adopted bin-based sampling to discard redundant voxels. Here we provide a performance comparison between our bin-based sampling and random sampling. The results are shown in Table 4. By adopting the same discarding rate of 90%, our bin-based sampling outperforms the random sampling by 3.14% AP, as our method can retain useful shape clues from faraway points.

More results on NuScenes test set. We have compared our method with CenterPoint + VP (virtual point), TransFusion-L + VP and TransFusion. The results on the nuScenes test set are shown in Table 5. With VirConv, the detection performance of CenterPoint + VP and TransFusion-L + VP has been significantly improved. Besides, the TransFusion-L with VirConv even surpasses the TransFusion in term of NDS, showing the effectiveness of our design. Our best results are also available on the evalAI online leaderboard (NuScenes evaluation server).

4 Visualization Results

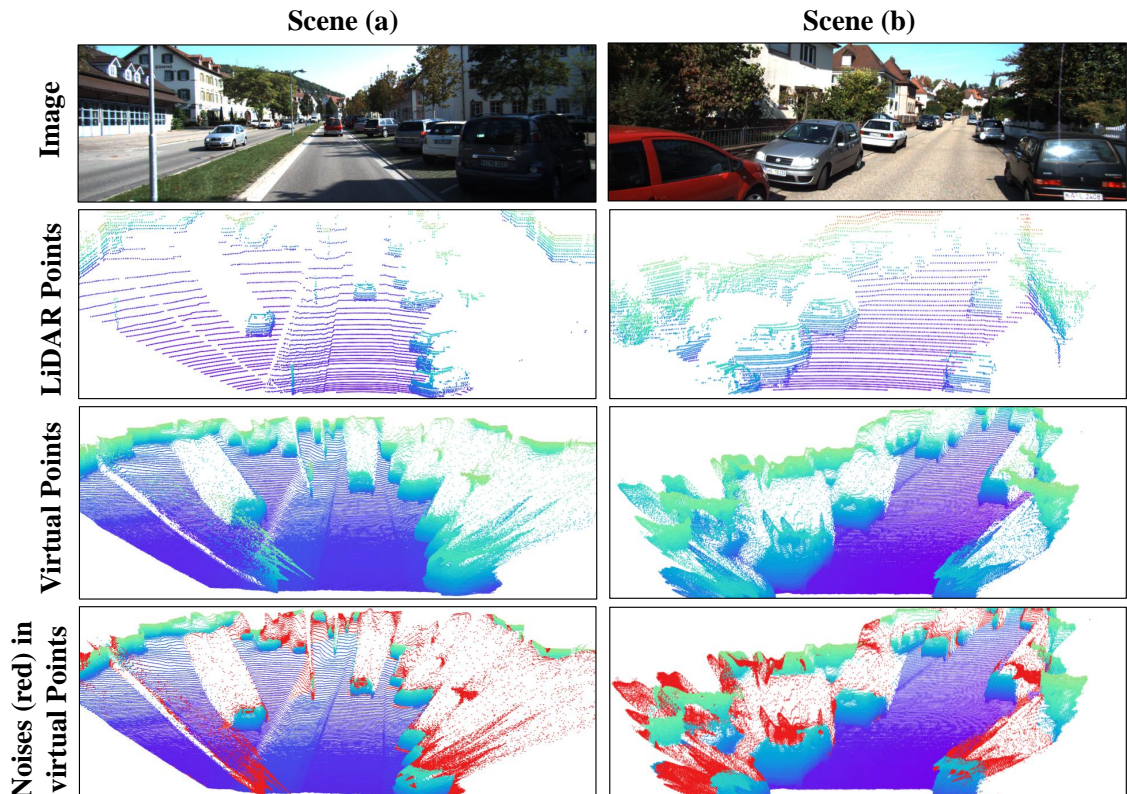


Figure 4: More examples of noisy and dense virtual points. We show two different scenes in (a) (b) respectively. The first, second and third rows show the RGB image, LiDAR points and virtual points, respectively. In the fourth row, we highlight the noises of virtual points in red color.

More examples of virtual points. To better understand the density and noise problem of virtual points, we provided more examples generated by PENet [2] in Fig. 4. We show two different scenes in (a - b), respectively. We observe that lots of noisy virtual points are distributed on the objects’ boundaries. The noises break the spatial structure of the object, bringing a significant challenge for accurate object localization. Besides, the virtual points are also much denser than regular LiDAR points. Our paper addressed these problems by NRConv and StVD, respectively.

Detection results. To better understand how our method improves the 3D detection performance, we show the visualization results of detected bounding boxes from the KITTI validation set in Fig. 5. For better comparison, we calculated the 3D IoU between ground truths and detections. We showed the comparison results of our VirConv-L and baseline (Voxel-RCNN [1]) in Fig. 5 (a)(c). We showed the comparison results of our VirConv-T and baseline (Voxel-RCNN [1]) in Fig. 5 (b)(d). We observe that our detected bounding boxes have higher 3D IoU with ground truth boxes, as our method can better leverage the geometry clue from virtual points to boost localization accuracy. Consequently, our models attain better detection performance.

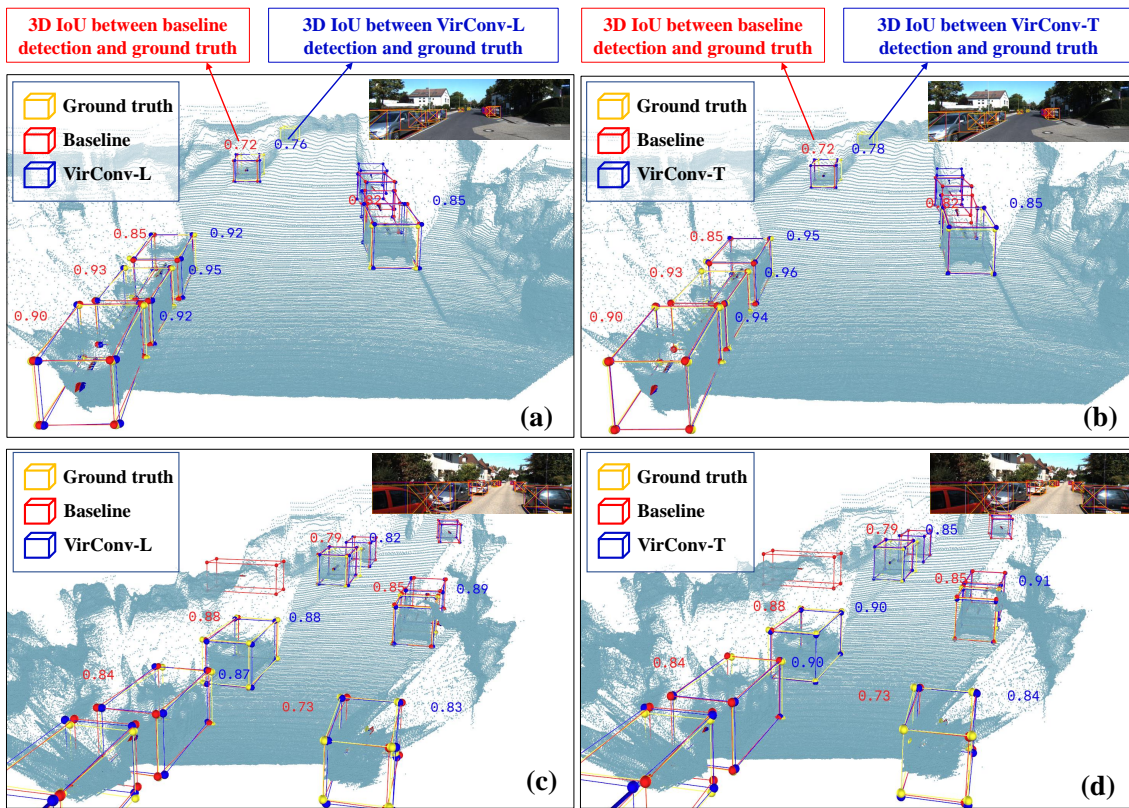


Figure 5: Visualization of detection results on KITTI validation set.

References

- [1] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wen gang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 6
- [2] Mu Hu, Shuling Wang, Bin Li, Shiyu Ning, Li Fan, and Xiaojin Gong. Penet: Towards precise and efficient image guided depth completion. In *International Conference on Robotics and Automation (ICRA)*, pages 13656–13662, 2021. 6
- [3] Hai Wu, Wenkai Han, Chenglu Wen, Xin Li, and Cheng Wang. 3d multi-object tracking in point clouds based on prediction confidence-guided data association. *IEEE TITS*, 2021. 3