Semi-Supervised Stereo-based 3D Object Detection via Cross-View Consensus Supplementary Material

Wenhao Wu[1], Hau-San Wong[1]*, and Si Wu[2]
[1]Department of Computer Science, City University of Hong Kong
[2]School of Computer Science and Engineering, South China University of Technology
wenhaowu5-c@my.cityu.edu.hk, cshswong@cityu.edu.hk, cswusi@scut.edu.cn

1. Overview

In the supplementary material, we present more experimental results and analysis as follows:

- We conduct an error analysis experiment to explore the possible reasons for the performance drop of the base model trained on limited annotated data only.
- We conduct an experiment to compare performances between the base models with the Left-Right (LR) disparity consistency constraint proposed by Godard *et al*. [2] and with our proposed Temporal-Aggregation-Guided (TAG) disparity consistency constraint.
- We conduct an ablation study to compare performances between the base models trained on different annotation ratios over the car and pedestrian categories on the KITTI validation set under different evaluation metrics.
- We provide a performance comparison between our proposed method and competing methods on the KITTI test set.
- We qualitatively compare detection results between the base model trained with and without pseudo-annotated data generated from our proposed method.

In this supplementary material, all experiments implemented on the KITTI validation set and test set are evaluated with 40 recall points, as the KITTI benchmark, for Average Precision (AP) calculation on the three modes of easy, moderate and hard.

2. Error Analysis of the Limited Supervision Setting

Inspired by CenterNet [11] and MonoDLE [5], we conduct an error analysis to explore the possible causes of performance drop due to limited supervision. In the error analysis, we replace the main attributes of predicted boxes generated from the base model trained on annotated data with those generated from the base model trained on fully-annotated data to determine the main restrictions of limited annotations. The results are shown in Tab. 1. We can observe that the refinement of the dimension and orientation attributes leads to a slight improvement in the detection performances. The refinement of the projected 3D center can further enhance the base model with a performance gain of 2.1% on AP_{3D}, since it contributes to better localization in the 3D space. However, the depth attribute contributes the most to the improvement of detection performances on both AP_{BEV} and AP_{3D}, verifying that the improvement of depth estimation can benefit the detection results to a significant extent when only limited annotated data are provided. Our proposed TAG method can stabilize and enhance the disparity estimation of the base model by the guidance of the more reliable and precise disparity maps from the teacher model with cumulative knowledge of the base model, thus contributing to further detection performance improvement as the results of the error analysis indicate.

3. Comparison between TAG-based and LR-based Disparity Consistency

We conduct an experiment to compare the base model trained with TAG-based and LR-based disparity consistency constraints. Specifically, we replace the disparity consistency constraint from our proposed TAG method with the LR method, in which we impose consistency between the output disparity maps of one view and the output disparity maps translated from the opposite view in the student model, while keeping other strategies, including data augmentation and the proposed cross-view agreement, unchanged for a fair comparison. The error comparison of the depth estimation is shown in Fig. 1. The direct consistency constraint between different views on the student

*Corresponding author.

Methods	AP_{BEV} (IoU=0.7)	Improvement	AP_{3D} (IoU=0.7)	Improvement
Baseline	27.93	-	18.09	-
+ sup. dimension	30.37	2.44	19.41	1.32
+ sup. orientation	30.47	2.54	19.40	1.31
+ sup. proj. center	29.47	1.54	20.19	2.1
+ sup. depth	36.38	**8.45**	24.40	**6.31**
Full Supervision	55.82	-	46.96	-

Table 1. Error analysis of the base model with limited supervision under the difficulty of moderate. We replace the attributes of 3D detection outputs from the base model trained on limited supervision with that from the base model trained on sufficient supervision to analyze the possible reasons that contribute to the performance drop when training on limited annotated data.

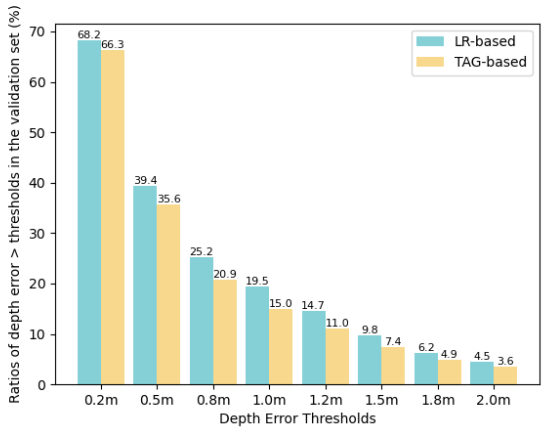

Figure 1. Error comparison of depth estimation between the base models with the LR-based disparity consistency constraint and our proposed TAG-based disparity consistency constraint when training on 5% of fully-annotated stereo images.

model results in poor depth estimation with more error in different depth ranges due to unreliable disparity outputs caused by limited supervision. The detection results are shown in Tab. 2. We can observe that the base model with the LR-based disparity consistency constraint, even under the same condition of data augmentation and pseudo annotation selection strategy, performs less satisfactorily compared to the base model with our proposed TAG-based disparity consistency constraint among all evaluation metrics, indicating that the LR-based disparity consistency hinders further improvement in 3D detection results.

4. Ablation Study of Annotation Ratio

We conduct an ablation study to analyze the cases when different ratios of annotated data are provided in the car category in Tab. 3 and in the pedestrian category in Tab. 4 under different evaluation metrics. In the car category, the base model with our proposed method achieves significant performance gains among different annotation ratios. Specifically, the base model with our proposed method achieves performance gains of up to 6%, 6% and 2% when 10%, 20% and 50% of fully-annotated data are provided, respectively. Similarly, our proposed method also enhances the separate base detector's performance in the pedestrian category even when the amount of pedestrian instances is significantly less than that of car instances, verifying the effectiveness of our proposed method in leveraging unannotated data to improve the detection performance of the base detector even in the category with a very small number of annotated instances.

5. 3D Object Detection on the KITTI test set

We report the performance comparison results between competing methods and our proposed method on the KITTI test set in Tab. 5. To further highlight the effectiveness of our proposed TAG method and cross-view agreement strategy, we leverage additional 8k unannotated data from the KITTI-raw dataset to enhance the base detector. The base model trained on both fully-annotated data and extra unannotated data leads to further performance improvement compared to the base model trained on fully-annotated data only, further highlighting the effectiveness of our proposed method in utilizing unannotated stereo images for achieving improved detection performance.

6. Qualitative Results

We visualize the 3D detection results of the base model, which is trained on 5% of full annotations, with and without pseudo-annotated data in both the front view and bird's eye view on the KITTI validation set, which is shown in Fig. 2. The base model trained on limited annotated stereo images performs poorly on the localization of distant objects, leading to unsatisfactory detection results. However, with our proposed TAG-based disparity consistency constraint, remote objects, which also pose significant difficulties for the LiDAR-based methods due to the very sparse LiDAR points, can be effectively detected for improving the overall

Methods	AP_{BEV}/AP_{3D} (IoU=0.5)			AP_{BEV}/AP_{3D} (IoU=0.7)		
	Easy	Mod.	Hard	Easy	Mod.	Hard
LR-based	83.56/79.58	57.38/53.63	43.91/40.64	47.93/31.47	29.87/19.29	22.67/14.68
TAG-based	87.18/83.48	66.39/60.85	50.36/47.07	53.89/35.91	36.30/24.50	27.99/18.81

Table 2. Performance comparison of average precision on bird's eye view (AP_{BEV}) and 3D boxes (AP_{3D}) between the base models, trained on 5% of fully-annotated stereo images, with the LR-based disparity consistency constraint and our proposed TAG-based disparity consistency constraint in the car category on the KITTI validation set.

Ratios	Methods	AP_{BEV}/AP_{3D} (IoU=0.5)			AP_{BEV}/AP_{3D} (IoU=0.7)		
		Easy	Mod.	Hard	Easy	Mod.	Hard
10%	Baseline	88.47/86.84	66.48/62.70	52.13/48.77	56.57/39.10	37.25/24.32	28.01/18.84
	Ours	91.84/88.30	72.63/67.22	56.11/52.83	63.07/44.48	43.58/30.84	33.47/22.73
20%	Baseline	90.02/89.50	70.00/68.90	59.76/52.27	72.69/60.00	47.86/38.13	39.06/30.38
	Ours	96.30/93.55	74.55/71.56	60.59/56.78	75.61/61.59	49.45/41.25	39.96/31.38
50%	Baseline	96.39/95.91	76.57/73.56	59.81/56.81	78.39/69.81	52.37/43.29	40.78/32.65
	Ours	96.94/96.25	77.08/74.13	62.13/59.14	78.71/70.26	54.36/44.89	41.29/33.11

Table 3. Performance comparison of average precision on bird's eye view (AP_{BEV}) and 3D boxes (AP_{3D}) between the base models trained on different ratios of annotated stereo images in the car category on the KITTI validation set.

performance. In addition, the base model with our proposed cross-view agreement strategy can detect some objects that are heavily occluded in the front view, verifying the effectiveness of the proposed cross-view agreement strategy in discovering occluded objects when provided only with stereo images.

References

[1] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Dsgn: Deep stereo geometry network for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12536–12545, 2020. 4

[2] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 1

[3] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7644–7652, 2019. 4

[4] Yuxuan Liu, Lujia Wang, and Ming Liu. Yolostereo3d: A step back to 2d for efficient stereo 3d detection. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13018–13024. IEEE, 2021. 4

[5] 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

[6] Xidong Peng, Xinge Zhu, Tai Wang, and Yuexin Ma. Side: Center-based stereo 3d detector with structure-aware instance depth estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 119–128, 2022. 4

[7] Jiaming Sun, Linghao Chen, Yiming Xie, Siyu Zhang, Qinhong Jiang, Xiaowei Zhou, and Hujun Bao. Disp r-cnn: Stereo 3d object detection via shape prior guided instance disparity estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10548–10557, 2020. 4

[8] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019. 4

[9] Zhenbo Xu, Wei Zhang, Xiaoqing Ye, Xiao Tan, Wei Yang, Shilei Wen, Errui Ding, Ajin Meng, and Liusheng Huang. Zoomnet: Part-aware adaptive zooming neural network for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12557–12564, 2020. 4

[10] Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, and Kilian Q. Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. In *International Conference on Learning Representations*, 2020. 4

[11] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 1

Ratios	Methods	AP_{BEV}/AP_{3D} (IoU=0.5)			AP_{BEV}/AP_{3D} (IoU=0.7)		
		Easy	Mod.	Hard	Easy	Mod.	Hard
10%	Baseline	26.33/26.29	19.51/19.40	15.89/15.79	12.24/10.91	8.79/7.54	6.90/5.83
	Ours	30.97/30.33	22.94/22.28	18.27/18.04	14.10/11.09	10.22/8.22	8.01/6.55
20%	Baselines	42.24/41.62	34.33/33.63	28.79/28.15	26.53/22.51	20.26/17.37	17.27/15.55
	Ours	44.71/43.15	36.97/36.01	30.81/30.14	27.35/23.11	21.12/18.02	18.18/16.11

Table 4. Performance comparison of average precision on bird's eye view (AP_{BEV}) and 3D boxes (AP_{3D}) between the base models trained on different ratios of annotated stereo images in the pedestrian category on the KITTI validation set.

Methods	Depth	AP_{3D} (IoU=0.7)			Time
		Easy	Mod.	Hard	(ms)
Stereo R-CNN [3]		47.58	30.23	23.72	280
SIDE [6]		47.69	30.82	25.68	260
ZoomNet [9]	✓	55.98	38.64	30.97	-
Disp R-CNN [7]	✓	59.58	39.34	31.99	425
Pseudo-LiDAR [8]	✓	54.53	34.05	28.25	670
Pseudo-LiDAR++ [10]	✓	61.11	42.43	36.99	510
DSGN [1]	✓	73.50	52.18	45.14	682
YoloStereo3D [4]	✓	65.68	41.25	30.42	160
Ours w/ unannot.	✓	68.00	43.33	31.04	160

Table 5. Performance comparison of average precision on 3D boxes (AP_{3D}) between our proposed method and competing methods, trained on fully-annotated data, on the KITTI test set. "Time" means inference time on the test set.