Unsupervised Domain Adaptation for Monocular 3D Object Detection via Self-Training

Zhenyu Li[1], Zehui Chen[2], Ang Li[3], Liangji Fang[3], Qinhong Jiang[3], Xianming Liu[1], and Junjun Jiang[1]✉

[1] Harbin Institute of Technology
[2] University of Science and Technology
[3] SenseTime Research
{zhenyuli17, csxm, jiangjunjun}@hit.edu.cn lovesnow@mail.ustc.edu.cn
{liang1, fangliangji, jiangqinhong}@senseauto.com

1 Dataset Comparisons

To provide more intuitive comparisons among different datasets (*e.g.*, KITTI [3], nuScense [1] and Lyft [4]), we present images with projected ground-truth labels in Fig. 1. One can easily observe cameras utilized in these datasets have different parameters, which are reflected in the image resolutions, FOV, *etc.* This work focuses on designing a general Mono3D UDA framework and solving the severe depth-shift caused by misaligned camera intrinsic parameters, which is the most crucial problem in Mono3D UDA. However, there are still numerous unsolved issues such as different image color styles, various distributions of object dimensions, different distributions of object depth, *etc.* Our proposed STMono3D can be a well-developed baseline for future research.

2 Visualizations of Pseudo Labels

Here, we present more visualizations of pseudo labels generated by the teacher model during the training stage. The images show the depth-shift issue caused by the misalignment of camera parameters can be well-solved. The reasonable pseudo labels provide regular supervision on the target domain and achieve Mono3D UDA in a teacher-student paradigm. In addition, we can find there is still a slight error of prediction locations or dimensions that can be improved by further development of Mono3D methods and enhancement of the UDA algorithms. There is still tremedous room for improvement of the Mono3D UDA.

3 Detailed Training Settings

In this section, we introduce more detailed training settings. As for the model, we follow the basic config provided in MMDetection3D [2].The only modification lies in the scaling of predicted object depth based on the pixel size (GAMS introduced in our paper). We then summary all the runtime settings in Tab. 1, including the number of interations, training schedule, threshold changing, *etc.*

Fig. 1. Dataset visualizations with ground-truth labels.

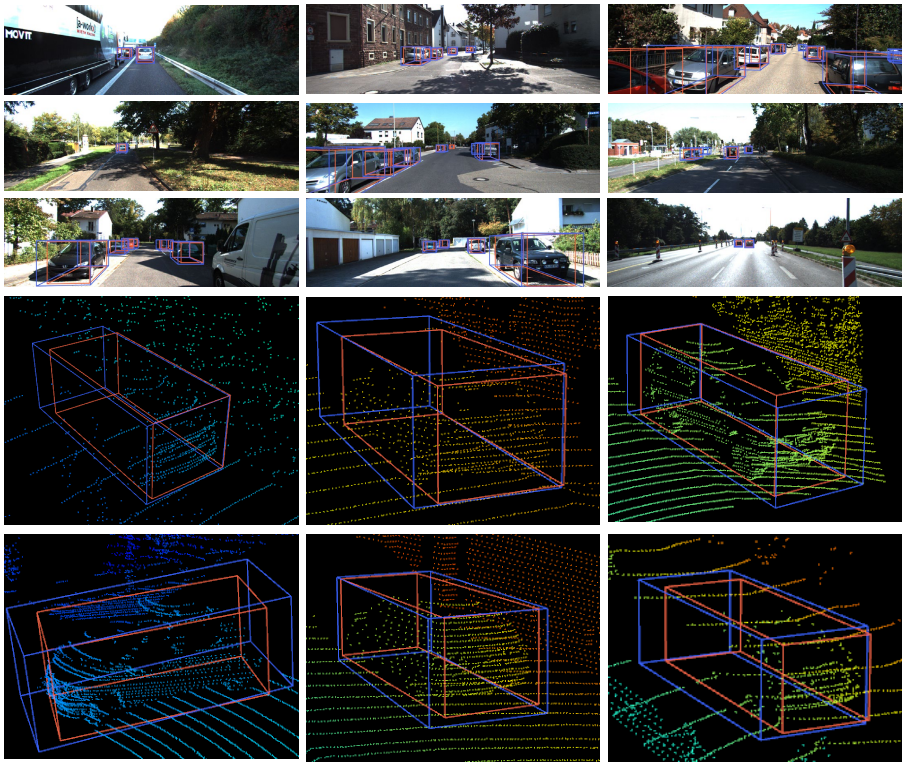

Fig. 2. Visualizations of pseudo labels generated by the teacher model.

References

1. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Computer Vision and Pattern Recognition (CVPR). pp. 11621–11631 (2020)
2. Contributors, M.: MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d (2020)
3. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Computer Vision and Pattern Recognition (CVPR). pp. 3354–3361 (2012)
4. Kesten, R., Usman, M., Houston, J., Pandya, T., Nadhamuni, K., Ferreira, A., Yuan, M., Low, B., Jain, A., Ondruska, P., Omari, S., Shah, S., Kulkarni, A., Kazakova, A., Tao, C., Platinsky, L., Jiang, W., Shet, V.: Level 5 perception dataset 2020. https://level-5.global/level5/data/ (2019)

Table 1. Detailed training settings.

Schedule	
number of iters	880 × 13
learning rate policy	step
warmup type	linear
warmup iters	500
warmup ratio	1/3
step	[880 × 8, 880 × 11]

Optimizer	
optimizer type	SGD
learning rate	0.002
gradient clip	True
batchsize per GPU	4
number of GPUs	8
source:domain samples per bs	1:1

Mean Teacher Paradigm	
momentum	0.999
interval	1
warmup	0
increasing thr per step	0.005
start iter	880×8
stop iter	880×10

Inference Settings (KITTI)		
student	nms pre	100
	nms thr	0.05
	score thr	0.001
	max per img	20
teacher	score thr (only diff.)	0.35

Strong Data Augmentation (type/prob/details)

RandomFlip3D	0.5	horizontal
Mono3DResize	1	(1600, 840) (1600, 900) (1600, 960) (1600, 1020) (1600, 1080) (1600, 1140) (1600, 1200) (1600, 1260) (1540, 840) (1480, 780) (1420, 720) (1380, 680) (1660, 960) (1720, 1020) (1800, 1080) (1880, 1140)
OneOf	1	Identity AutoContrast RandEqualize RandSolarize RandColor RandContrast RandBrightness RandSharpness RandPosterize
RandErase	1	size=[0, 0.2] n blocks=(1, 5) squared=True

Weak Data Augmentation (type/prob/details)

RandomFlip3D	0.5	horizontal
Mono3DResize	1	(1600, 840) (1600, 900) (1600, 960) (1600, 1020) (1600, 1080) (1600, 1140) (1600, 1200) (1600, 1260) (1540, 840) (1480, 780) (1420, 720) (1380, 680) (1660, 960) (1720, 1020) (1800, 1080) (1880, 1140)