DetMatch: Two Teachers are Better Than One for Joint 2D and 3D Semi-Supervised Object Detection (Supplementary Materials)

Jinhyung Park[1*], Chenfeng Xu[2†], Yiyang Zhou[2], Masayoshi Tomizuka[2], and Wei Zhan[2]

[1] Carnegie Mellon University, Pittsburgh PA 15213, USA
jinhyun1@andrew.cmu.edu
[2] University of California, Berkeley, Berkeley CA 94720, USA
{xuchenfeng, yiyang.zhou, tomizuka, wzhan}@berkeley.edu

In this supplementary, we provide additional training details, quantitative results, and qualitative visualizations. These sections are organized as follows:

- Section 1 provides additional details about our implementation and training setup.
- Section 2 contains the ablation study for the thresholds of the single-modality 2D and 3D SSL pipelines.
- Section 3 contains the ablation study for performing 2D NMS on projected 3D boxes before 2D-3D matching.
- Section 4 presents additional quantitative results with DetMatch using SECOND instead of PV-RCNN, demonstrating the adaptability of our method.
- Section 5 lists training and inference costs of our method.
- Section 6 contains additional visualizations of DetMatch on KITTI and Waymo.

1 Additional Implementation and Training Details

For all settings, including single-modality pipelines and DetMatch, we pre-train on the labeled data, then initialize both the teacher and student model with these weights to begin SSL training. Similar to prior work [11,8], we ramp-up EMA momentum from 0.99 to 0.999. We follow Unbiased Teacher [3] in only supervising classification for the 2D detector in both single-modality SSL and DetMatch. Further, we adopt their setting of maintaining a constant learning rate through training and taking the teacher as the final model - we find that this yields more stable and reproducible performance. Moreover, we replace the classification loss in the 2D detector with Focal Loss [2], which has been found [3] to yield more class-balanced pseudo-labels in SSL. The 3D detectors PV-RCNN [6] and SECOND [9] use Focal Loss by default, so they are left unchanged. For fair

* Work conducted during visit to University of California, Berkeley.
† Corresponding author

comparison, we use the same 2D weak-strong augmentations as Unbiased Teacher and the same 3D weak-strong augmentations as 3DIoUMatch. All experiments were conducted on three NVIDIA A6000 GPUs.

1.1 KITTI Training Details

Pre-training. Owing to the smaller number of samples, we pre-train the 2D detector for 240 epochs and the 3D detector for 800 epochs on the 1%, 2% settings and for 120 and 400 epochs on the 20% setting. An epoch is defined as one pass over the limited labeled data, and we find that these long cycles allow for full convergence. The batch size is 24 for both 2D and 3D detectors. The 2D detector is trained using SGD with starting learning rate 0.03 decayed 10x twice mimicking the standard "1x" COCO training cycle. PV-RCNN is trained using AdamW [4] with the one-cycle scheduling strategy and a max learning rate of 0.015. SECOND is trained with the same settings and PV-RCNN but with a max learning rate of 0.0045

SSL Training. Each SSL batch consists of 12 labeled and 12 unlabeled samples, and we use the starting or max learning rate of pre-training as the constant learning rate when training 2D or 3D SSL. Based on observations in other multi-modality works [10], we use a separate optimizers for 2D and 3D in DetMatch, maintaining SGD for 2D and AdamW for 3D.

1.2 Waymo Training Details

Pre-training. We use the Waymo v1.0 released data. We pre-train the 2D detector for a standard "1x" COCO training cycle on the 1% setting and train the 3D detector for 48 epochs. Due to the much higher resolution of Waymo's 2D images, we do not do multiscale training, instead keeping the original 1920x1280 resolution. Additionally, due to GPU memory limitations, a single 2D sample in Waymo consists of the front view and a side view, the latter sampled from one of the four side images. We do this because the front view has far more objects on average than the other views. Finally, the detectors are trained with half the batch size and learning rate as they were in KITTI.

SSL Training. Each SSL batch consists of 6 labeled and 6 unlabeled samples. We find that for Waymo, the raw LiDAR intensity value wildly varies from 0 to tens of thousands, causing instability in the early layers. As such, we freeze the first block for 3D detectors during SSL training. For DetMatch, the 2D teacher predicts boxes on all five 2D views for 2D-3D teacher Hungarian Matching, but only two 2D views are used to train the 2D student due to memory limitations. Since the five 2D images have a combined FOV of 240 degrees, we simply use confidence thresholding on the 3D teacher to generate pseudo-labels on the remaining 120 degrees. Despite not being able to apply DetMatch on the full 3D scene, we find that our pipeline improves over the 3D SSL baseline.

Table 1: Impact of τ_{3D}

3D Eval	mAP	Car	Ped	Cyc
Labeled-Only	45.9	73.8	30.4	33.4
$\tau_{3d} = 0.2$	48.6	75.2	33.4	37.1
$\tau_{3d} = 0.3$	**54.4**	75.9	**42.7**	**44.6**
$\tau_{3d} = 0.4$	50.6	**76.4**	35.0	40.3
$\tau_{3d} = 0.5$	45.7	72.7	31.4	42.7

Table 2: Impact of τ_{2D}

2D Eval	mAP	Car	Ped	Cyc
Labeled-Only	65.3	86.6	68.6	40.8
$\tau_{3d} = 0.6$	55.7	84.1	67.6	15.4
$\tau_{3d} = 0.7$	**60.4**	86.1	**69.2**	**25.8**
$\tau_{3d} = 0.8$	57.5	**88.0**	60.4	24.3

Table 3: Ablation on 2D NMS for projected 3D boxes

1% Data	3D			2D		
	Car	Ped	Cyc	Car	Ped	Cyc
Without 2D NMS	77.4	55.8	42.7	88.5	72.1	51.4
With 2D NMS	77.5	57.3	42.3	88.8	73.9	51.7

2 Thresholds for 2D and 3D Single-Modality SSL

We extensively search for the best confidence thresholds τ_{3D}, τ_{2D} for our 3D and 2D SSL baselines. The results are shown in Tables 1 and 2. We observe that for both modalities, although the best mAP is acheived at $\tau_{3D} = 0.3$ and $\tau_{2D} = 0.7$, Car detection peaks at a slightly higher threshold. 3DIoUMatch adopts different thresholds for Car, but to avoid introducing additional hyperparameters, we use a single threshold for both SSL baselines and DetMatch. It is worth noting, however, that even if we had used a class-specific threshold for Car for the single-modality SSL baselines, our DetMatch still achieves better Car performance.

For 2D-only SSL, we exhaustively searched confidence thresholds, training schedules, and weighting parameters but were unable to find a setting that yields improved performance on all classes for KITTI 1%. As mentioned in the main paper, this can be attributed to the more difficult and limited-data setting of SSL on autonomous driving datasets as well as the single-modality self-training error propagation. Indeed, we find that on KITTI 20% results shown in the main paper, 2D SSL is able to improve performance, demonstrating that more labeled data is required to improve on the already-strong 2D labeled-only baseline.

3 2D NMS before 2D-3D Matching

As 3D boxes lose depth information when projected to 2D, we find that noisy 3D false positive boxes occupying the same image-frustum as a true positive 3D box crowd the correct 3D box in the image, sometimes leading to incorrect 2D-3D matches. We carefully address this problem by doing 2D NMS on the projected 3D boxes before 2D-3D matching. In Table 3, we find that this benefits pedestrians, which are most likely to have spurious 3D detections, the most.

Table 4: KITTI Results for DetMatch + SECOND

1% Data	3D				2D			
	mAP	Car	Ped	Cyc	mAP	Car	Ped	Cyc
Labeled-Only	38.3	65.4	22.6	26.9	65.3	86.6	68.6	40.8
Confidence Thresholding	38.8	70.1	26.7	19.7	60.4	86.1	69.2	25.8
Improvement	+0.5	+4.7	+4.1	-7.2	-4.9	-0.5	+0.6	-15.0
SESS [11]	38.9	64.5	29.4	22.6	-	-	-	-
Improvement	+0.6	-0.9	+6.8	-4.3	-	-	-	-
Ours	49.4	74.9	41.9	31.5	68.5	88.7	70.3	46.5
Improvement	**+11.1**	**+9.5**	**+19.3**	**+4.6**	**+3.2**	**+2.1**	**+1.7**	**+5.7**

Table 5: Waymo Results for DetMatch + SECOND

1% Data	3D								2D			
	Car L1		Car L2		Ped L1		Ped L2		Car		Ped	
	mAP	mAPH	mAP	mAPH	mAP	mAPH	mAP	mAPH	L1	L2	L1	L2
Labeled-Only	35.6	34.4	32.6	31.5	19.7	10.4	17.8	9.4	42.3	39.5	50.8	47.0
Confidence Thresholding	42.7	41.8	40.1	39.3	27.7	13.3	25.1	12.1	44.4	41.3	48.7	45.1
Improvement	+7.1	+7.4	+7.5	+7.8	+8.0	+2.9	+7.3	+2.7	+2.1	+1.8	-2.1	-1.9
Ours	45.2	44.1	41.5	40.6	35.7	16.9	32.3	15.3	48.1	44.8	51.1	47.1
Improvement	**+9.6**	**+9.7**	**+8.9**	**+9.1**	**+16.0**	**+6.5**	**+14.5**	**+5.9**	**+5.8**	**+5.3**	**+0.3**	**+0.1**

4 DetMatch with SECOND

To demonstrate the adaptability of DetMatch, we replace the two-stage PV-RCNN 3D detector with a representative one-stage 3D detector SECOND. The KITTI results are shown in Table 4 and the Waymo results are in Table 5.

For additional SOTA comparison for 3D SSL, we also try to adapt SESS [11], a 3D SSL method developed for VoteNet [5], an indoor one-stage model, to autonomous driving. Despite extensive searches, we were unable to improve on labeled-only PV-RCNN [6], an outdoor two-stage model, with SESS. However, we compare with SESS on SECOND [9], a one-stage model in Table 4.

We find that on KITTI, although confidence thresholding is able to substantially improve 3D Car and Pedestrian results, it reduces performance for Cyclist. Since Cyclist is a rare category, with only a dozen samples in the KITTI 1% setting, we find that SECOND, a weaker but substantially faster 3D detector than PV-RCNN, is not able to sufficiently learn this class to generate accurate pseudo-labels for self-training. A similar trend can be observed in SESS's results, except SESS does better on Pedestrians but worse on Cars. Our DetMatch addresses this problem, more accurately identifying high quality pseudo-labels by considering consistency between 2D and 3D detections. DetMatch substantially improves performance in all metrics. Further, unlike other approaches [8,1], DetMatch does not require additional modules to estimate box quality, being readily adaptable to various detectors.

We see a similar trend on the Waymo dataset as shown in Table 5. Our DetMatch improves on both the labeled-only model and the strong single-modality SSL baselines, notably improving 3D Pedestrian performance by 16.0 mAP and 2D Car performance by 5.7 mAP. Interestingly, unlike DetMatch with PV-

Table 6: Training costs with PV-RCNN + Faster-RCNN R50

Method	KITTI		Waymo	
	GPU Mem. (GB)	GPU Hrs	GPU Mem. (GB)	GPU Hrs
2D Conf. Thresh.	18.5	4.1	25.8	71.0
3D Conf. Thresh.	22.2	8.4	13.5	29.5
Ours	38.3	11.2	37.9	75.0

RCNN, we find that DetMatch with SECOND improves significantly over the 3D-only SSL baseline for the Car class. We attribute this to SECOND being a weaker detector than PV-RCNN and thus being able to derive more benefits from joint training with the 2D detector. Overall, we find that DetMatch is adaptable, able to easily work with various detectors, and that its single hyperparameter τ_{hung} is robust under various settings. This is due τ_{hung} thresholding on a *consistency* cost between 2D and 3D detections, which is fundamentally different from simple confidence thresholding. Predicted class confidence is a single model's evaluation of its own predictions, subject to self-bias and error propagation through training. On the other hand, consistency cost can be considered one model "consulting" another to evaluate its predictions. As this cost depends on *agreement* between semantic class predictions and box parameters of these two models, it is a measurement of box quality more decoupled from any single model. Drawing from these benefits, our DetMatch demonstrates improvement over labeled-only and single-modality SSL methods.

5 Training & Inference Costs

5.1 Training Costs

In Table 6, we detail training costs under settings in our supplementary. Surprisingly, DetMatch costs less than the 2D and 3D baselines together.

Alongside GPU-side optimizations, this is because most of DetMatch's additions, such as Hungarian Matching, are CPU operations that do not bottleneck throughput. Waymo's 2D baseline GPU hours are high because loading Waymo's high-res images imposes an I/O bottleneck. Finally, DetMatch's single-sample batch GPU memory usage for KITTI and Waymo are 9.2GB and 19.3GB respectively, allowing training on consumer 2080Ti and 3090 GPUs.

5.2 Inference Costs

We emphasize that DetMatch is a framework for *training* generic 2D & 3D detection models, leaving their architectures unmodified. As such, DetMatch imposes **no additional inference-time burdens** beyond those of the chosen 2D and 3D detection models themselves. For completeness, we still report Faster-RCNN R50 + FPN's and PV-RCNN's inference times on our setup with an A6000 GPU. Faster-RCNN's inference times on KITTI and Waymo are 31.4 FPS and 2.2 FPS respectively. PV-RCNN's inference times on KITTI and Waymo

are 10.5 FPS and 2.5 FPS respectively. Please note that our machine has a significant I/O bottleneck loading 5 high-res images for Waymo, which causes the low 2.2 FPS for Faster-RCNN. However, please again note that these are the unadulterated inference speeds of the single-modality detection models; our DetMatch leaves them unchanged, hence not impacting their inference times. If more stringent FPS constraints must be met, one can use DetMatch with lighter 2D or 3D detectors, as demonstrated in Section 4 with the lightweight SECOND 3D detector.

6 Qualitative Results

We present qualitative results of DetMatch for KITTI and Waymo in Figures 1 and 2, respectively. Note that although point cloud is colored for visualization, 3D-only detectors take as input a color-less point cloud so that they can be applied to LiDAR-only setups as well.

We find that our method is effective in utilizing the advantages of 2D detections to preserve correct 3D detections and vice versa. 2D detections are especially useful in removing false positive 3D detections and "promoting" accurate but low confidence pedestrian 3D detections that are filtered away in 3D-only SSL. On the other hand, we observe that 3D detections are better at objects highly occluded in 2D because these objects are clearly separated in 3D. This property allows DetMatch to generate clear and accurate 2D boxes even for highly overlapped cars. We also notice some error cases of DetMatch. When an object correctly detected in one modality does not have a corresponding prediction in the other modality, the correct detection is not preserved as a pseudo-label. Although this causes our method to miss some objects, DetMatch generates far fewer false positives than single-modality SSL pipelines, and many works with very high thresholds [7,12] have observed that such a set of precise, albeit sparser, pseudo-labels is preferable to many noisy labels. We will investigate how to leverage objects correctly detected in only one modality in future work.

Fig. 1: Visualizations of DetMatch on KITTI

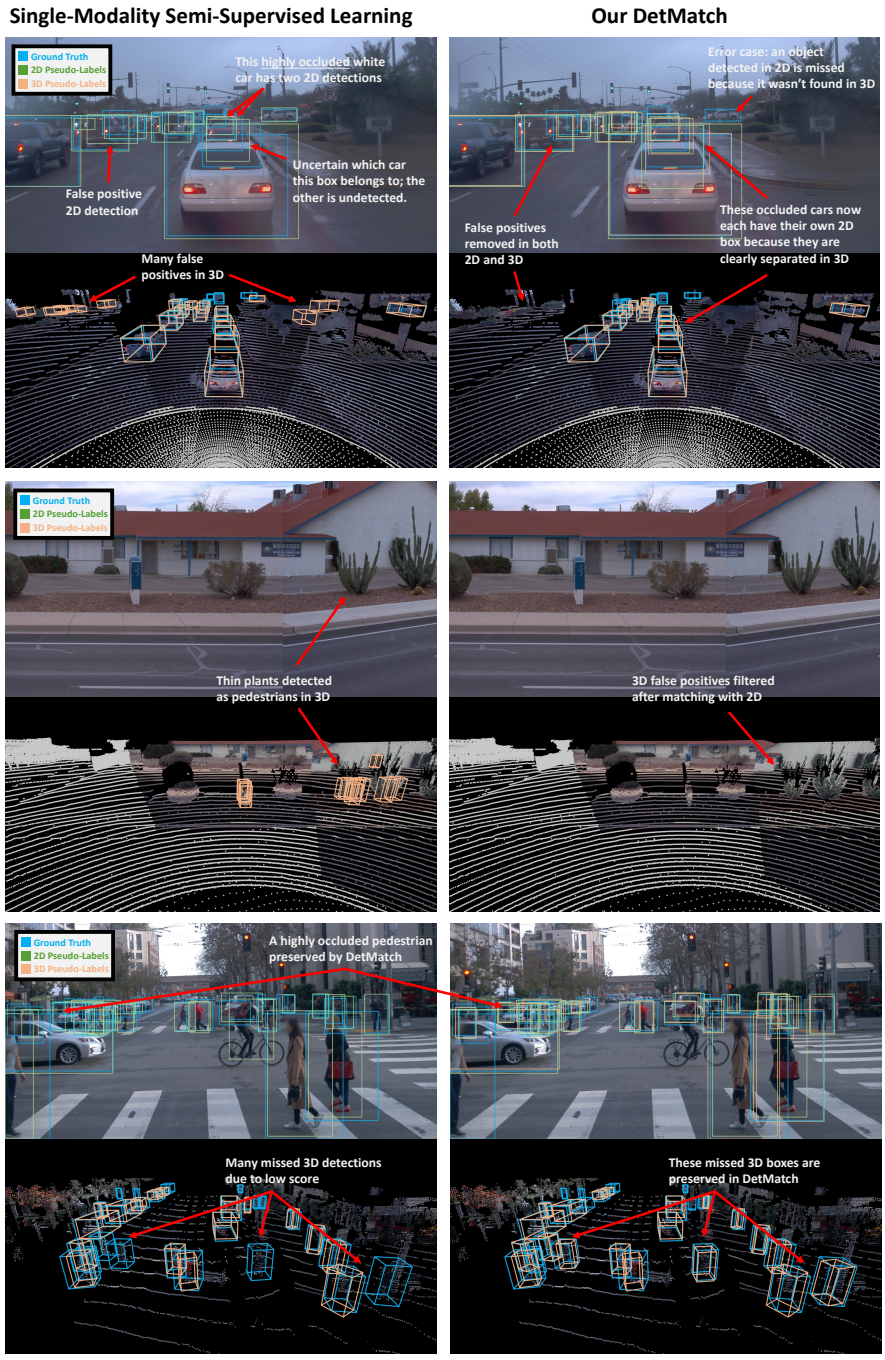

Fig. 2: Visualizations of DetMatch on the Waymo Dataset

References

1. Li, H., Wu, Z., Shrivastava, A., Davis, L.S.: Rethinking pseudo labels for semi-supervised object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1314–1322 (2022)
2. Lin, T.Y., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**, 318–327 (2020)
3. Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. ICLR (2021)
4. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
5. Qi, C., Chen, X., Litany, O., Guibas, L.: Imvotenet: Boosting 3d object detection in point clouds with image votes. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4403–4412 (2020)
6. Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H.: Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10526–10535 (2020)
7. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in Neural Information Processing Systems **33**, 596–608 (2020)
8. Wang, H., Cong, Y., Litany, O., Gao, Y., Guibas, L.J.: 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 14610–14619 (2021)
9. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors (Basel, Switzerland) **18** (2018)
10. Zhang, W., Wang, Z., Loy, C.C.: Multi-modality cut and paste for 3d object detection. ArXiv **abs/2012.12741** (2020)
11. Zhao, N., Chua, T.S., Lee, G.H.: Sess: Self-ensembling semi-supervised 3d object detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 11076–11084 (2020)
12. feng Zhou, Q., Yu, C., Wang, Z., Qian, Q., Li, H.: Instant-teaching: An end-to-end semi-supervised object detection framework. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4079–4088 (2021)