Leveraging Temporal Cues for Semi-Supervised Multi-View 3D Object Detection (Supplementary Materials)

Jinhyung Park[1] Navyata Sanghvi[1] Hiroki Adachi[1] Yoshihisa Shibata[2] Shawn Hunt[3]
Shinya Tanaka[3] Hironobu Fujiyoshi[2] Kris Kitani[1]
[1]Carnegie Mellon University [2]DENSO Corporation [3]DENSO International America, Inc.

A. Supplementary Overview

In this supplementary, we provide additional implementation details, analysis on forwards vs backwards detection performance, and more qualitative results. The sections are organized as follows:

- Section B shows additional results on the Argoverse 2 benchmark.
- Section C.1 provides implementation details regarding the training of our framework.
- Section D includes evaluation results on forward vs backward detections.
- Section E contains an ablation on the pseudo-labeling thresholds.
- Section F presents additional comparisons between confidence pseudo-labeling and our framework. We also include a supplementary video visualizing these pseudo-labels.
- Section G shows visualizations of RGB reconstruction results from our model. We also include a supplementary video.

B. Performance on Argoverse 2

In Table 6, we further evaluate our framework on Argoverse 2 [7], using 16k labeled and 94k unlabeled frames. Argoverse 2 contains 26 different classes, and we limit detection to a 102.4m x 102.4m x 8m range centered on the ego-vehicle, similar to nuScenes and nuPlan. Our framework demonstrates consistent improvements.

C. Implementation Details

C.1. Details of Our Framework

We build on the strong, temporal 3D detector StreamPETR [6] and adopt the ConvNeXt-S [4] backbone pre-trained by SparK [5]. Our framework is trained jointly with detection losses and the RGB reconstruction loss. To avoid biasing the detector to the masked-input distribution, we first run the backbone and the detection head on uncorrupted RGB images, then run the backbone on the images masked with a ratio of 0.3. We supervise the masked patches using a z-normed reconstruction target. The object query-conditioned masked reconstruction uses two transformer decoder layers, allowing each image feature location to extract features from object queries to aid the reconstruction.

During training, we evenly sample sequences from each geographic location — Las Vegas, Boston, Pittsburgh, and Singapore for NuPlan [2], and Boston and Singapore for NuScenes [1]. We find this approach maintains performance in all locations, especially relevant for NuPlan which has most of its data from Las Vegas. To avoid training instability from suddenly adjusting the learning rate schedule between the first stage (labeled) and the second stage (labeled and unlabeled) training, we maintain a fixed, high learning rate for the model and maintain the exponential moving average (EMA) of its weights. Using this approach, the EMA model is used for pseudo-labeling, and the base model maintains the same learning rate over training stages, allowing for more stable optimization. Our final loss consists of StreamPETR's detection losses (3D detection, auxiliary 2D detection) and our L2 reconstruction loss. The weight for the reconstruction loss is set as 1. We use a learning rate of 3.75e-4 for a batch size of 15 in the first stage, and equally sample 15 frames from both labeled and unlabeled data in the second stage [3]. Following StreamPETR, we use a resized image size of 256 x 704. We evaluate our model on the full nuScenes validation set. For NuPlan, we randomly select a subset of 10k frames from the official validation split for faster evaluation. We emphasize that such a 10k subset is already much larger than the nuScenes validation set.

C.2. Details of Baselines

For the Pseudo-Labeling baseline, we generate labels with a confidence threshold of 0.4. The training settings are identical between the baseline and our framework. On nuScenes, to establish a strong baseline building on UniPAD, we first pre-train UniPAD on the entire nuScenes dataset and transfer the image backbone to StreamPETR. We then fine-tune this detector using the limited labels and further refine the

Method	Front of Ego-Vehicle			Back of Ego-Vehicle			All Objects		
	mAP ↑	NDS ↑	mATE ↓	mAP ↑	NDS ↑	mATE ↓	mAP ↑	NDS ↑	mATE ↓
Forwards Inference	0.170	0.251	0.881	0.203	0.269	0.866	0.188	0.261	0.870
Backwards Inference	0.185	0.261	0.864	0.176	0.249	0.916	0.183	0.257	0.886
FwBw Ensembling	0.185	0.261	0.864	0.203	0.269	0.866	**0.196**	**0.266**	**0.863**

Table 5. **Performance of methods on objects in front vs. behind the ego-vehicle, and across all objects.** The detector is trained with 800 labeled samples and with forwards-backwards sampling during training.

Method	mAP ↑	CDS ↑	mATE ↓	mASE ↓	mAOE ↓
Labeled Only	0.129	0.088	0.938	0.382	0.722
Pseudo-Labeling	0.150	0.104	0.993	0.397	0.857
Ours	**0.158**	**0.110**	0.933	0.394	0.884

Table 6. Comparison on the Argoverse 2 dataset.

Thr.	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓
-0.5	0.216	0.276	0.857	0.162	1.114	1.197
-1.0	**0.219**	**0.278**	0.850	0.162	1.102	1.200
-1.5	0.215	0.277	0.847	0.163	1.100	1.190

Table 7. Ablation of Hungarian Matching threshold on nuPlan.

baseline by pseudo-labeling the unlabeled data and continuing training. Consequently, the UniPAD + pseudo-labeling method replaces the pseudo-labeling baseline as the primary comparison target on nuScenes.

D. Forward vs Backwards Performance

In Table 5 we evaluate a 3D detector trained on 800 labeled frames on objects in-front of the ego-vehicle and behind the ego-vehicle. When the detector is run forwards in time, we observe it performs substantially better on objects behind the vehicle. On the other hand, when the detector is run backwards, the model improves performance on objects ahead of the ego-vehicle while compromising performance on those behind. By ensembling predictions from both forwards and backwards inference of the *same detector,* we achieve the best of both worlds and substantially improve overall pseudo-label quality.

E. Ablation on Pseudo-Labeling Threshold

We ablate the hungarian matching threshold for pseudo-labeling for our method in Table 7. This experiment uses with 800 labeled and 60k unlabeled frames on nuPlan, and we choose -1.0. We find that direct pseudo-labeling is robust to the threshold – 0.3, 0.4, 0.5 perform similarly, and we choose 0.4 for its slightly better convergence.

F. Qualitative Analysis of Pseudo-Labels

In Figures 5 and 6, we visualize the pseudo-labels of a detector trained on 800 labeled samples. Ground truth is shown in green, the confidence pseudo-labeling method is shown in orange, and our framework is in blue. Even in this low-cost setting, our pipeline yields more coherent and consistent pseudo-labels, in turn substantially improving the performance of the final model trained on these pseudo-labels.

G. Qualitative Analysis of RGB Reconstruction

In Figure 7 we show pairs of masked input and reconstructed output. Even when using a comparatively lightweight backbone (ConvNeXt-S with 50M parameters vs SparK's flagship ConvNeXt-L results with 200M parameters), our model generates plausible reconstructions of masked regions. The model leverages object query conditioning to fill in even almost entirely occluded objects.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1

[2] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021. 1

[3] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021. 1

[4] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 1

[5] Keyu Tian, Yi Jiang, Qishuai Diao, Chen Lin, Liwei Wang, and Zehuan Yuan. Designing bert for convolutional networks:

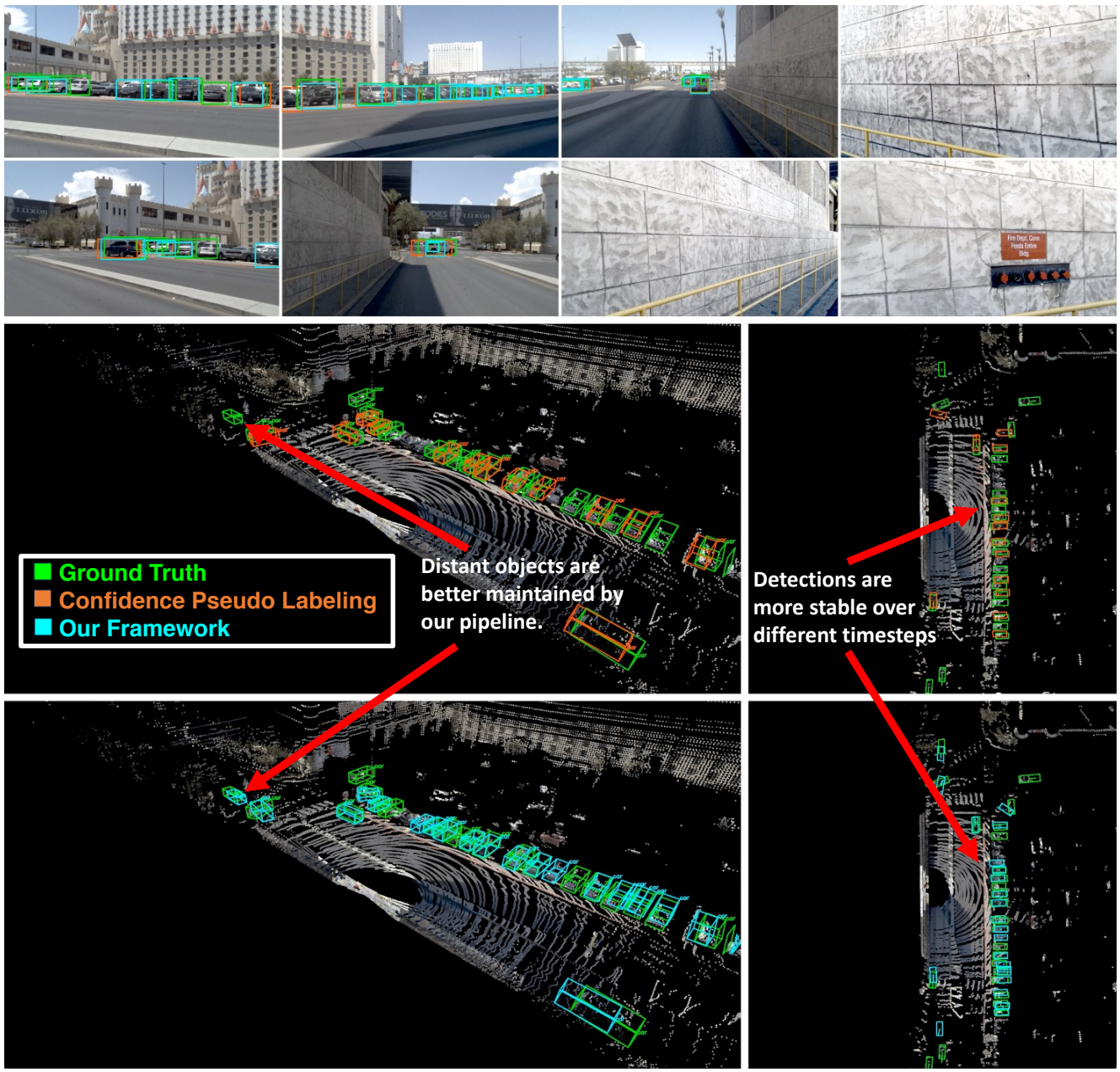

Figure 5. **Qualitative results on pseudo-labeling.** By fully leveraging temporal cues, our framework yields more consistent pseudo-labels for both close and far objects.

Sparse and hierarchical masked modeling. *arXiv:2301.03580*, 2023. 1

[6] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3621–3631, 2023. 1

[7] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 1

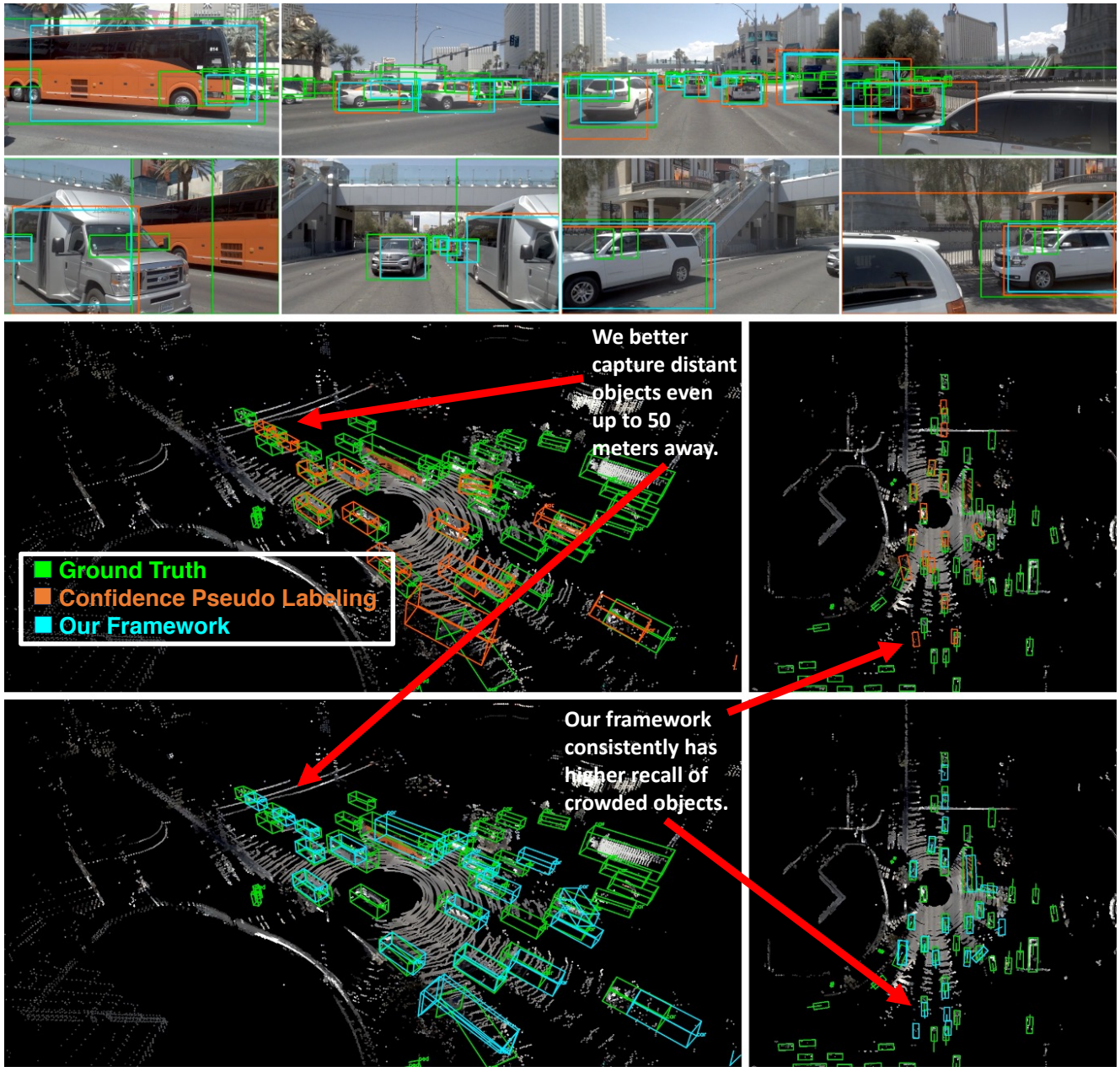

Figure 6. **Qualitative results on pseudo-labeling.** Even in a difficult and crowded driving scenario, our pipeline better captures highly occluded objects.

Figure 7. **Qualitative results of RGB Reconstruction.** We show pairs of masked images and predicted reconstructions. The model learns the overall shape and appearance of objects, generating plausible reconstructions even when objects are largely occluded.