

# SOAP: Cross-sensor Domain Adaptation for 3D Object Detection Using Stationary Object Aggregation Pseudo-labelling

Chengjie Huang    Vahdat Abdelzad    Sean Sedwards    Krzysztof Czarnecki  
 University of Waterloo

{c.huang, vahdat.abdelzad, sean.sedwards, k2czarne}@uwaterloo.ca

## Abstract

We consider the problem of cross-sensor domain adaptation in the context of LiDAR-based 3D object detection and propose Stationary Object Aggregation Pseudo-labelling (SOAP) to generate high quality pseudo-labels for stationary objects. In contrast to the current state-of-the-art in-domain practice of aggregating just a few input scans, SOAP aggregates entire sequences of point clouds at the input level to reduce the sensor domain gap. Then, by means of what we call quasi-stationary training and spatial consistency post-processing, the SOAP model generates accurate pseudo-labels for stationary objects, closing a minimum of 30.3% domain gap compared to few-frame detectors. Our results also show that state-of-the-art domain adaptation approaches can achieve even greater performance in combination with SOAP, in both the unsupervised and semi-supervised settings.

## 1. Introduction

LiDAR sensors are commonly used in autonomous driving and other safety-critical robotic applications to provide accurate 3D localization of objects. State-of-the-art (SOTA) LiDAR-based object detectors currently use deep neural networks trained via supervised learning, requiring a large amount of realistic data labelled by human annotators. Annotation is expensive, motivating the re-use of existing labelled datasets, but these typically have a limited *domain*, e.g., a specific sensor configuration and relatively few geographic locations and weather conditions. It is well known that detectors trained with a dataset from one domain experience a significant degradation of performance when fed data from a different domain.

In this work, we tackle the cross-sensor domain adaptation problem that arises whenever a detector is required to interpret data from a sensor different to that on which it was trained. Specifically, given a detector trained on labelled point clouds collected using one sensor (the *source*

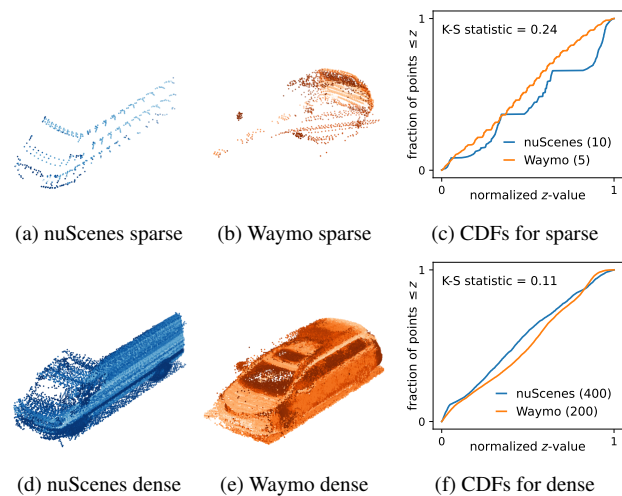


Figure 1. Scan lines are evident in point clouds when only few input frames are used (a)(b), appearing as obvious modes in CDF plots that largely differ because of the modes (c). Aggregating many more frames removes visible scan lines (d)(e) and makes CDFs for similar objects in different datasets more alike (f)

*domain*), we aim to improve the detector’s performance on point clouds collected using a different sensor (the *target domain*), either with no labels (unsupervised) or with only a small number of labels from the target domain (semi-supervised). This situation arises commonly when a LiDAR sensor is updated or a fleet of autonomous vehicles uses multiple sensors.

Cross-sensor domain adaptation can present a formidable challenge because similar objects scanned by different LiDAR sensors may have very different scan patterns, even after the widely-adopted few-frame input aggregation. Figures 1a and 1b show point clouds of vehicles at similar distances from their respective sensors, created by aggregating 10 frames from nuScenes [2] and 5 frames from Waymo [16] datasets, respectively. There are evident

scan lines and obvious visual dissimilarities between the point clouds. There are also significant differences in the corresponding cumulative density function (CDF) plots of the  $z$ -component of point positions (Fig. 1c).

Current SOTA 3D object detectors, while achieving impressive in-domain performance, still have a considerable performance gap when they are applied to cross-sensor point clouds, whether using single- or few-frame input. This is demonstrated in Table 1, where we compare the performance of a VoxelNeXt [4] model trained and evaluated on the nuScenes [2] and Waymo [16] datasets.

We attribute this drop in performance in large part to the different scan patterns mentioned above. This view is supported by a recent study based on simulation [8] that suggests the difference in scan patterns alone can have a substantial impact on the performance of existing object detectors. At the same time, we observe that aggregating many more frames tends to reduce the scan patterns. This is illustrated in Figs. 1d and 1e, which show dense point clouds created by aggregating 400 nuScenes and 200 Waymo frames, respectively. There are no visible scan lines and the multi-modality evident with few frames has disappeared from the CDF plots in Fig. 1f.

SOTA methods in cross-sensor domain adaptation often employ pseudo-labelling, where a model trained on labelled data is used to generate labels for unlabelled data. Various approaches have been proposed to improve pseudo-label quality and regularize training with pseudo-labels [21, 25, 26], but these approaches do not appear to explicitly address the important difference in scan-patterns between different domains.

Given all of the above, we propose *Stationary Object Aggregation Pseudo-labelling* (SOAP) to improve cross-sensor pseudo-label accuracy by exploiting scene-level full-sequence aggregation of input point clouds to close the domain gap caused by scan-patterns.

SOAP uses sequential point clouds produced under realistic driving conditions by existing LiDAR sensors that are widely used in autonomous driving systems. It enhances existing pre-trained detectors, improving their stationary object performance while retaining dynamic object pseudo-labels. SOAP pseudo-labels can be used for updating detectors or bootstrapping annotations.

SOAP is motivated by the facts that (i) aggregation improves the representation of sparsely-scanned objects [19], (ii) sensor-specific scan patterns are reduced by full-sequence aggregation (Fig. 1), and (iii) stationary objects respond well to full-sequence aggregation and are a statistically important component of object detection: at least two thirds of cars are stationary at some point in sequences in major realistic driving datasets [2, 10, 16, 22].

Extensive experiments using nuScenes and Waymo datasets show SOAP pseudo-labels can close a minimum

|                              | # frames | AP   | $\Delta$     |
|------------------------------|----------|------|--------------|
| nuScenes $\rightarrow$ Waymo | 1        | 7.0  | <b>-70.5</b> |
|                              | 5        | 20.4 | <b>-57.1</b> |
| Waymo in-domain              | 5        | 77.5 |              |
| Waymo $\rightarrow$ nuScenes | 1        | 42.7 | <b>-40.8</b> |
|                              | 10       | 49.0 | <b>-34.5</b> |
| nuScenes in-domain           | 10       | 83.5 |              |

Table 1. Degradation ( $\Delta$ ) w.r.t. in-domain performance of SOTA detector VoxelNeXt [4] in Waymo  $\leftrightarrow$  nuScenes cross-sensor setting. Increasing the number of aggregated frames improves over single-frame input, but there is still a substantial performance gap.

of 30.3% overall domain gap compared to few-frame detectors without access to any target domain labels. SOAP also complements other SOTA domain adaptation methods, including ST3D [26] and SSDA3D [21], improving their already impressive results in both unsupervised and semi-supervised settings. In nuScenes  $\rightarrow$  Waymo setting using CenterPoint [29], for example, using SOAP closes 42.6% domain gap compared to the 9.5% closed by ST3D. With only 1% target domain labels, SOAP closes 86.8% domain gap compared to the 81.4% closed by SSDA3D.

Our main contributions are as follows:

- We propose SOAP to effectively utilize full-sequence scene-level aggregation and exploit the properties of the pseudo-labels.
- We demonstrate that full-sequence scene-level aggregation, though not optimal for in-domain settings, can be used to improve cross-sensor performance.
- We conduct extensive experiments to demonstrate SOAP’s high quality pseudo-labels and synergy with SOTA domain adaptation methods.

## 2. Related Work

**Domain adaptation for 3D object detection:** Many methods have been proposed to address the domain adaptation problem for 3D object detection. One line of work involves improving model robustness via regularization [19, 23]. Other works attempt to close the domain gap via domain mapping [1, 5] or input [17, 18, 20], feature [12, 28, 31], and output [6] alignment. To take advantage of available target domain data, SOTA methods often use pseudo-labelling. Pseudo-labels can be improved via tracking-based refinement [9, 30] or iterative self-training [25, 26]. When a small amount of target labels are available, CutMix and MixUp have also been shown to be effective techniques to incorporate labelled target data [21]. As we will show, SOAP is parallel to and can complement existing work in this area.

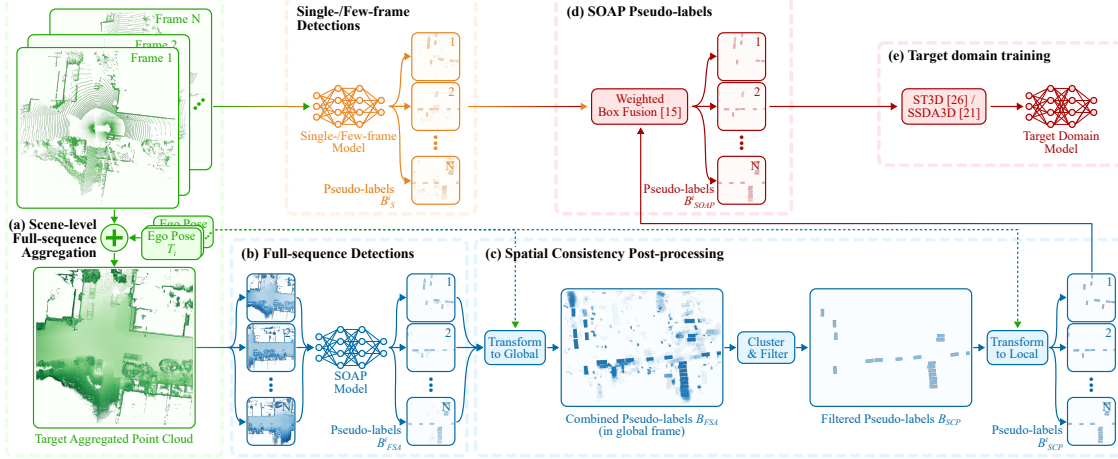


Figure 2. Overview of Stationary Object Aggregation Pseudo-labelling (SOAP) pipeline. (a) We first perform Full-Sequence Aggregation (FSA) using pose transforms. (b) We propose Quasi-Stationary Training (QST) to train a SOAP model to detect stationary objects. (c) The predictions are refined via Spatial Consistency Post-processing (SCP). (d) The predictions from a pre-trained single-/few-frame detector and the SOAP model are combined using Weighted Box Fusion (WBF) [15]. (e) The final SOAP pseudo-labels can be used in combination with SOTA methods to fine-tune a target domain detector.

**Offline pseudo-labelling:** Previous studies have shown that increasing the number of frames aggregated at scene-level leads to diminishing returns [2] or even performance degradation [3], especially for dynamic objects [27]. As a result, SOTA offline pseudo-labelling methods use a single- or few-frame detector to generate initial predictions, followed by offline tracking and a second stage refinement that utilizes full-sequence point clouds aggregated at object- or track-level [7, 13, 14, 24]. Offline pseudo-labelling has achieved impressive in-domain results, even surpassing human performance, but they have not yet been explored in cross-sensor setting. SOAP, on the other hand, takes a completely different view from the aforementioned works and directly uses scene-level aggregated point clouds as input to provide better pseudo-labels than single- or few-frame detectors for cross-sensor domain adaptation setting.

### 3. Our approach: SOAP

In this section, we describe the details of *Stationary Object Aggregation Pseudo-labelling* (SOAP). The main components of SOAP are: (i) *Scene-level Full-sequence Aggregation* (SFA), which produces aggregated point clouds from the entire input sequence; (ii) *Quasi-Stationary Training* (QST) that is used to train a pseudo-labelling model to detect stationary objects; and (iii) *Spatial Consistency Post-processing* (SCP) that enhances pseudo-labels by exploiting the stationarity of the predictions. SOAP is used in combination with a pre-trained model to generate high quality pseudo-labels for stationary objects while retaining dynamic object pseudo-labels. An overview of the SOAP pipeline is shown in Fig. 2.

### 3.1. Scene-level Full-sequence Aggregation

*Scene-level full-sequence aggregation* (SFA) involves projecting a sequence of point clouds to a global coordinate system, where the point clouds are concatenated into a single dense point cloud, as illustrated in Fig. 2a. Formally, given a sequence of point clouds  $P_1, P_2, \dots, P_N$ , where  $P_i = \{p_i^1, p_i^2, \dots, p_i^{M_i}\} \subset \mathbb{R}^3$ , and corresponding sequence of  $\mathbb{SE}(3)$  pose transformations  $T_1, T_2, \dots, T_N$ , which transform the point clouds from the local LiDAR or vehicle coordinate system to a common global coordinate system, the point cloud aggregation process in the global coordinate system is defined by

$$P^* = \bigcup_{i=1,2,\dots,N} \{T_i P_i^j\}_{j=1,2,\dots,M_i}. \quad (1)$$

During training, for a given frame  $i$ , the aggregated point cloud  $P^*$  is transformed back to the local coordinate system using the inverse pose transform  $T_i^{-1}$ .

Scene-level aggregation of a few frames in a short time window has been shown to be effective at producing denser input point clouds and consequently better detection performance [2]. On the other hand, due to the diminishing returns [2] and performance degradation [3, 27] observed for scene-level aggregation with large temporal windows, full-sequence aggregation has only been attempted at object- and track-level [7, 13, 14, 24].

Compared to few-frame aggregation, SFA increases point density and provides richer geometric information for stationary objects. In addition, unlike object- and track-level aggregation used by prior work, SFA does not rely on object annotations or initial predictions produced by single-

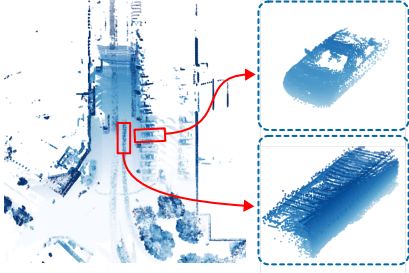


Figure 3. Example of a point cloud generated by FSA. Dynamic objects are distorted while stationary objects are densified.

or few-frame detectors, which we show are unreliable in the cross-sensor setting Table 1. SFA is therefore very suitable for its proposed application, since  $P^*$  can be obtained for both the source and (unlabelled) target domains.

SFA tends to distort dynamic object point clouds, due to the motion of the objects not being corrected during aggregation. This is depicted in Fig. 3. By contrast, stationary objects are densified with a more complete and accurate geometry compared to single- or few-frame point clouds. As noted above and in Fig. 1, the aggregation process also weakens the LiDAR-specific scan patterns.

### 3.2. Quasi-Stationary Training

Although annotations are available for the source domain, training a model to detect stationary objects from aggregated point clouds is not straightforward. A naive approach is by filtering the ground truth annotations during training based on speed estimates or the overall displacement of the object. However, since the speed of an object can change over time, especially for sequences that span a large time window, the object can be near stationary during part of the sequence and moving in other parts. If the majority of the observed points come from the part of the sequence where the object is near stationary, the object’s aggregated point cloud will have little distortion, as if the object were stationary for the entirety of the sequence.

We refer to such objects as *quasi-stationary* objects. Figure 4 depicts an example of such an object, which would be excluded based on a naive speed or displacement criterion. Doing so would result in undistorted objects remaining in the aggregated point clouds, but without a positive label, causing confusion and reducing model performance.

To avoid excluding these quasi-stationary objects, we formally define the notion of quasi-stationarity using a *quasi-stationary score* (QSS) that takes into account both the movement of the objects and how much each observation contributes to the final aggregated point clouds.

**Definition 1 (QSS)** Let  $\{b_1, b_2, \dots, b_N\}$  be a set of  $N$  bounding boxes of an object annotated in a sequence within a common coordinate system and  $C(b_i)$  be the number of

points observed in the bounding box  $b_i$ . For a given bounding box observation  $b_i$ , QSS is defined as the average IoU between  $b_i$  and other bounding boxes  $b_j$ , weighted by the fractions of points contributed by  $b_j$ :

$$\text{QSS}(b_i) = \sum_{j=1}^N \frac{C(b_j)}{\sum_{k=1}^N C(b_k)} \text{IoU}(b_i, b_j) \quad (2)$$

Intuitively, the QSS can be interpreted as how likely it is that the point cloud for a given object is undistorted at the location of  $b_i$ . For example, if another observation  $b_j$  has little overlap with  $b_i$  (indicating object movement) but contains only a few points, then the final aggregated point cloud is not likely to be distorted by  $b_j$ . Alternatively, if  $b_j$  has a large overlap with  $b_i$  and also contains a large fraction of points, then the object is likely to be undistorted at location  $b_i$ .

Finally, the most likely location  $b^*$  of the object in the aggregated point clouds and the degree  $s^*$  of the point cloud being free from distortion can be estimated as follows:

$$b^* = \arg \max_i \text{QSS}(b_i) \quad (3)$$

$$s^* = \max_i \text{QSS}(b_i) \quad (4)$$

We refer to  $b^*$  as the quasi-stationary bounding box and  $s^*$  as the corresponding QSS. Objects with a large QSS  $s^* > \epsilon$  for some  $\epsilon$  can be considered quasi-stationary. For instance, the object in Figure 4 has QSS  $s^* = 0.91$ .

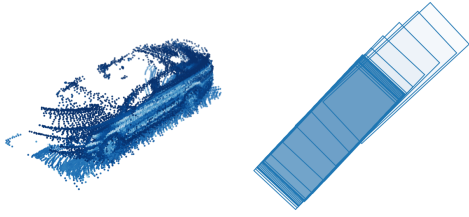
**Out-of-sight quasi-stationary objects:** In the labelled source domain dataset, we notice objects not visible from the current frame are sometimes not labelled. Since SFA utilizes the entire sequence to create a dense representation of the scene, out-of-sight quasi-stationary objects will also be densified in the corresponding aggregated point cloud. This creates inconsistent labels where in a given frame, some dense objects have annotations while others do not. To address this problem, we construct spatially consistent training labels by projecting the quasi-stationary bounding boxes  $b^*$  to *all* frames in the sequence, even if the object is not observed in some frames.

### 3.3. Spatial Consistency Post-processing

Since the SOAP model is tasked to detect only quasi-stationary objects from aggregated point clouds, a detected object should have consistent predictions in the global coordinate system across multiple frames. To utilize this stationarity property of the pseudo-labels, we propose *Spatial Consistency Post-processing* (SCP) to eliminate false positive predictions and recover false negative objects, improving pseudo-label quality.

As illustrated in Fig. 2c, SCP is performed by obtaining per-frame predicted bounding boxes (denoted by  $B_{FSA}^i$  for frame  $i$ ) using the SOAP model and then gathering





(a) Aggregated point cloud (b) Object trajectory (BEV)

Figure 4. Example of a quasi-stationary object. This object reached a maximum speed of 1.4 m/s with a total displacement of 3.9 m, and thus would be eliminated by naive filtering.

all predictions in the global coordinate system. Gathering all predictions requires transforming per-frame predictions (bounding boxes) using the corresponding ego pose transformation  $T_i$ :

$$B_{FSA} = \bigcup T_i B_{FSA}^i \quad (5)$$

The bounding boxes  $B_{FSA}$  are clustered based on an IoU threshold  $\mu$ . To ensure the pseudo-labels are consistent across multiple frames, we eliminate the clusters with a number of detections fewer than threshold  $\eta$  that depends on the frame rate of the dataset.

For the remaining clusters, the boxes in each cluster are combined into a single bounding box per cluster, similar to Weighted Boxes Fusion (WBF) [15]. Specifically, we use the heading  $\theta$  of the most confident prediction in each cluster and average other attributes, including position  $(x, y, z)$ , size  $(w, l, h)$  and velocity  $(v_x, v_y)$ , weighted by the confidence of each box. Finally, non-maximum suppression is applied to remove any overlapping predictions in the global coordinate system. The final bounding boxes in the global coordinate system are denoted by  $B_{SCP}$ .

To obtain the pseudo-labels  $B_{SCP}^i$  for each frame  $i$  in the sequence, we transform  $B_{SCP}$  back to each frame's local coordinate system using the inverse pose transformation  $T_i^{-1}$ , defined as follows:

$$B_{SCP}^i = T_i^{-1} B_{SCP} \quad (6)$$

Since an object may be occluded in a frame and have very few or no points in the sparse point cloud, we remove any bounding boxes that contain no points, so that the pseudo-labels are reasonable with respect to the frame's sparse point cloud.

### 3.4. SOAP pseudo-labels

In order to recover a complete set of pseudo-labels for both stationary and dynamic objects, SOAP utilizes the predictions from a pre-trained single- or few-frame detector capable of detecting dynamic objects. The SOAP and pre-trained models are calibrated separately with Beta Calibration [11] using source domain data. Let  $B_S^i$  be the bounding

boxes predicted by the pre-trained detector for frame  $i$ , then the SOAP pseudo-labels, denoted by  $B_{SOAP}^i$ , are obtained by combining  $B_S^i$  and  $B_{SCP}^i$  using WBF [15].

Our results show that this simple approach can improve existing sparse pseudo-labels by a large margin. We leave it as future work to study more optimal ways of combining pseudo-labels or obtaining dynamic pseudo-labels directly from aggregated point clouds.

## 4. Experiments

SOAP is evaluated in both unsupervised and semi-supervised domain adaptation settings. This section presents the experimental setup, domain adaptation results, and ablation study.

### 4.1. Datasets

We evaluate SOAP using two large-scale autonomous driving datasets for 3D object detection: nuScenes [2] and Waymo [16]. In what follows, we use the syntax *source domain dataset*  $\rightarrow$  *target domain dataset* to denote the training and testing setting, respectively.

The NuScenes dataset [2] contains 1,000 sequences of 20 seconds each, collected in Boston and Singapore. The vehicle is equipped with a single Velodyne HDL-32E 32-beam top-mounted rotating LiDAR operating at 20 Hz, yielding  $\approx 400$  point cloud scans per sequence, from which 40 keyframes are selected uniformly and annotated. Unless specified otherwise, all models trained on the nuScenes dataset use 10 sweeps ( $-0.5$  s) as input.

The Waymo dataset [16] contains 1,150 sequences of 20 seconds each, collected in San Francisco, Mountain View, and Phoenix. The Waymo dataset uses a 5-sensor setup with a single proprietary 64-beam top-mounted rotating LiDAR operating at 10 Hz, and four side-mounted close-range LiDARs. All point clouds ( $\approx 200$ ) in each sequence are annotated with bounding boxes. Due to much higher annotation frequency compared to nuScenes, we use 20% uniformly sampled frames for training. Consistent with the nuScenes models, all models trained on the Waymo dataset use 5 sweeps ( $-0.5$  s) as input.

In addition to the difference in the point cloud and annotation frequency, the nuScenes dataset annotates 23 classes with 8 attributes, 10 of which are used in the object detection task, whereas Waymo contains annotations for only Vehicle, Cyclist, and Pedestrian. Following previous work [12, 17, 18, 20, 25, 26, 28, 30], we select the common vehicle/car class for training and evaluation for all experiments.

### 4.2. Evaluation

For nuScenes evaluation, we consider two primary metrics: mean Average Precision (mAP) and NuScenes Detection Score (NDS). Following the official evaluation, mAP

| Architecture     | Method             | Training Data                    | Overall            |                    | 0–30 m             |                    | 30–50 m            |                    |
|------------------|--------------------|----------------------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
|                  |                    |                                  | Level 1            | Level 2            | Level 1            | Level 2            | Level 1            | Level 2            |
| CenterPoint [29] | Direct             | $\{\mathcal{S}\}$                | 23.5               | 20.2               | 49.3               | 48.3               | 12.0               | 10.5               |
|                  | SOAP (ours)        | $\{\mathcal{S}\}$                | <b>50.9</b> +51.0% | <b>45.4</b> +51.2% | <b>69.4</b> +47.0% | <b>68.4</b> +47.1% | <b>47.5</b> +55.5% | <b>43.6</b> +55.5% |
|                  | ST3D [26]          | $\{\mathcal{S}, \mathcal{T}_P\}$ | 28.6 +9.5%         | 24.6 +8.9%         | 56.5 +16.8%        | 55.3 +16.4%        | 18.7 +10.5%        | 16.5 +10.1%        |
|                  | ST3D + SOAP (ours) | $\{\mathcal{S}, \mathcal{T}_P\}$ | <b>46.4</b> +42.6% | <b>40.6</b> +41.5% | <b>68.6</b> +45.1% | <b>67.5</b> +45.0% | <b>41.8</b> +46.6% | <b>37.6</b> +45.5% |
|                  | Oracle             | $\{\mathcal{T}\}$                | 77.2               | 69.4               | 92.1               | 91.0               | 76.0               | 70.1               |
| VoxelNeXt [4]    | Direct             | $\{\mathcal{S}\}$                | 20.4               | 17.5               | 44.2               | 43.3               | 9.5                | 8.4                |
|                  | SOAP (ours)        | $\{\mathcal{S}\}$                | <b>50.9</b> +53.4% | <b>45.6</b> +53.8% | <b>67.5</b> +48.5% | <b>66.4</b> +48.3% | <b>48.6</b> +58.5% | <b>44.7</b> +58.6% |
|                  | ST3D [26]          | $\{\mathcal{S}, \mathcal{T}_P\}$ | 35.0 +25.6%        | 30.2 +25.3%        | 60.7 +34.4%        | 59.5 +33.9%        | 27.9 +27.5%        | 24.8 +26.5%        |
|                  | ST3D + SOAP (ours) | $\{\mathcal{S}, \mathcal{T}_P\}$ | <b>45.6</b> +44.1% | <b>40.0</b> +43.1% | <b>65.9</b> +45.2% | <b>64.8</b> +45.0% | <b>41.8</b> +48.4% | <b>37.7</b> +47.3% |
|                  | Oracle             | $\{\mathcal{T}\}$                | 77.5               | 69.7               | 92.2               | 91.1               | 76.3               | 70.3               |

$\mathcal{S}$ : labelled source domain;  $\mathcal{T}$ : labelled target domain;  $\mathcal{T}_P$ : pseudo-labelled target domain

Table 2. Unsupervised domain adaptation results for nuScenes  $\rightarrow$  Waymo, where Waymo dataset is unlabelled. The percentages represent the amount of the Direct–Oracle domain gap closed.

| Architecture     | Method             | Training Data                    | Overall            |                    | 0–30 m             |                    | 30–50 m            |                    |
|------------------|--------------------|----------------------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
|                  |                    |                                  | mAP                | NDS                | mAP                | NDS                | mAP                | NDS                |
| CenterPoint [29] | Direct             | $\{\mathcal{S}\}$                | 51.7               | 69.6               | 67.6               | 78.9               | 27.6               | 49.7               |
|                  | SOAP (ours)        | $\{\mathcal{S}\}$                | <b>61.4</b> +30.3% | <b>76.9</b> +39.2% | <b>73.1</b> +21.7% | <b>83.9</b> +33.3% | <b>41.9</b> +40.1% | <b>64.4</b> +60.2% |
|                  | ST3D [26]          | $\{\mathcal{S}, \mathcal{T}_P\}$ | 59.3 +23.8%        | 72.9 +17.7%        | <b>74.2</b> +26.1% | 82.4 +23.3%        | 34.1 +18.2%        | 52.7 +12.3%        |
|                  | ST3D + SOAP (ours) | $\{\mathcal{S}, \mathcal{T}_P\}$ | <b>61.5</b> +30.6% | <b>75.4</b> +31.2% | 73.9 +24.9%        | <b>83.0</b> +27.3% | <b>42.5</b> +41.7% | <b>60.4</b> +43.9% |
|                  | Oracle             | $\{\mathcal{T}\}$                | 83.7               | 88.2               | 92.9               | 93.9               | 63.3               | 74.1               |
| VoxelNeXt [4]    | Direct             | $\{\mathcal{S}\}$                | 49.0               | 68.3               | 62.5               | 76.5               | 28.8               | 50.6               |
|                  | SOAP (ours)        | $\{\mathcal{S}\}$                | <b>61.5</b> +36.2% | <b>77.0</b> +44.2% | <b>72.5</b> +32.5% | <b>83.5</b> +40.5% | <b>43.9</b> +45.5% | <b>65.3</b> +65.3% |
|                  | ST3D [26]          | $\{\mathcal{S}, \mathcal{T}_P\}$ | 54.2 +15.1%        | 70.8 +12.7%        | <b>64.6</b> +6.8%  | 77.8 +7.5%         | 38.0 +27.7%        | 55.3 +20.9%        |
|                  | ST3D + SOAP (ours) | $\{\mathcal{S}, \mathcal{T}_P\}$ | <b>56.0</b> +20.3% | <b>72.7</b> +22.3% | <b>64.6</b> +6.8%  | <b>78.3</b> +10.4% | <b>43.7</b> +44.9% | <b>61.2</b> +47.1% |
|                  | Oracle             | $\{\mathcal{T}\}$                | 83.5               | 88.0               | 93.3               | 93.8               | 62.0               | 73.1               |

$\mathcal{S}$ : labelled source domain;  $\mathcal{T}$ : labelled target domain;  $\mathcal{T}_P$ : pseudo-labelled target domain

Table 3. Unsupervised domain adaptation results for Waymo  $\rightarrow$  nuScenes, where nuScenes dataset is unlabelled. The percentages represent the amount of the Direct–Oracle domain gap closed.

is calculated based on four distance thresholds (0.5, 1.0, 2.0, 4.0) and averaged. As distance-based mAP does not penalize other types of bounding box errors, NDS is used in combination to reflect the average translation, scale, orientation, velocity, and attribute errors for the true positive predictions. All evaluations are performed on the validation split consisting of 150 sequences.

For Waymo evaluation, we use the official evaluation suite and report the Level 1 and Level 2 AP scores. Different from nuScenes mAP, Waymo AP is calculated based on 3D IoU with a threshold of 0.7. Level 1 evaluation includes only objects with more than 5 points within the bounding box, while Level 2 evaluation considers all objects. All evaluations are performed on the validation split consisting of 202 sequences.

### 4.3. Unsupervised domain adaptation

We first evaluate SOAP pseudo-labels in the unsupervised domain adaptation setting, where annotations from the target domain are unavailable. We compare SOAP with two baseline approaches: “Direct” and ST3D [26]. Direct is where a few-frame detector is trained on the source domain

and directly evaluated on the target domain data. ST3D is a SOTA unsupervised domain adaptation method based on pseudo-labelling and self-training.

In the baseline comparison, ST3D utilizes the Direct model to generate pseudo-labels for self-training. To demonstrate the quality of the SOAP pseudo-labels and the complementary nature of SOAP with other approaches, we further consider ST3D + SOAP, where ST3D uses SOAP pseudo-labels. Both ST3D experiments use the official code release.

SOAP is validated using two object detection architectures: CenterPoint [29] and VoxelNeXt [4]. CenterPoint is a widely-adopted voxel-based dense 3D object detector. VoxelNeXt is a SOTA architecture representing recent advances in fully-sparse 3D object detectors. Both architectures are based on the implementation in the open-source library OpenPCDet. We use the Direct model predictions as few-frame predictions to construct final SOAP pseudo-labels. More implementation detail and hyper-parameters can be found in the supplementary material. The main results are shown in Tables 2 and 3.

| Method               | Training Data                                   | Overall     |               | 0–30 m      |               | 30–50 m     |               |
|----------------------|---|-------------|---------------|-------------|---------------|-------------|---------------|
|                      |   | Level 1     | Level 2       | Level 1     | Level 2       | Level 1     | Level 2       |
| Direct               | $\{\mathcal{S}\}$                               | 23.5        | 20.2          | 49.3        | 48.3          | 12.0        | 10.5          |
| Co-training          | $\{\mathcal{S}, \mathcal{T}_C\}$                | 58.4        | +65.0%        | 51.3        | +63.2%        | 79.3        | +70.1%        |
| CutMix [21]          |   | 58.6        | +65.4%        | 51.6        | +63.8%        | 78.9        | +69.2%        |
| SOAP (ours)          |   | <b>69.4</b> | <b>+85.5%</b> | <b>63.1</b> | <b>+87.2%</b> | <b>84.6</b> | <b>+82.5%</b> |
| SSDA3D [21]          | $\{\mathcal{S}, \mathcal{T}_C, \mathcal{T}_P\}$ | 67.2        | +81.4%        | 59.8        | +80.5%        | 84.5        | +82.2%        |
| SSDA3D + SOAP (ours) |   | <b>70.1</b> | <b>+86.8%</b> | <b>62.6</b> | <b>+86.2%</b> | <b>86.1</b> | <b>+86.0%</b> |
| Oracle               | $\{\mathcal{T}\}$                               | 77.2        | 69.4          | 92.1        | 91.0          | 76.0        | 70.1          |

$\mathcal{S}$ : labelled source domain;  $\mathcal{T}$ : labelled target domain;  $\mathcal{T}_C$ : small subset of  $\mathcal{T}$ ;  $\mathcal{T}_P$ : pseudo-labelled target domain

Table 4. Semi-supervised domain adaptation results for nuScenes  $\rightarrow$  Waymo, where 1% of Waymo data is labelled. The percentages represent the amount of the Direct–Oracle domain gap closed.

| Method               | Training Data                                   | Overall     |               | 0–30 m      |               | 30–50 m     |               |
|----------------------|---|-------------|---------------|-------------|---------------|-------------|---------------|
|                      |   | mAP         | NDS           | mAP         | NDS           | mAP         | NDS           |
| Direct               | $\{\mathcal{S}\}$                               | 51.7        | 69.6          | 67.6        | 78.9          | 27.6        | 49.7          |
| Co-training          | $\{\mathcal{S}, \mathcal{T}_C\}$                | 64.1        | +38.8%        | 76.4        | +36.6%        | 80.3        | +50.2%        |
| CutMix [21]          |   | 63.6        | +37.2%        | 76.7        | +38.2%        | 82.3        | +58.1%        |
| SOAP (ours)          |   | <b>74.0</b> | <b>+69.7%</b> | <b>83.5</b> | <b>+74.7%</b> | <b>86.4</b> | <b>+74.3%</b> |
| SSDA3D [21]          | $\{\mathcal{S}, \mathcal{T}_C, \mathcal{T}_P\}$ | 72.7        | +65.6%        | 81.3        | +62.9%        | 88.1        | +81.0%        |
| SSDA3D + SOAP (ours) |   | <b>76.8</b> | <b>+78.4%</b> | <b>83.1</b> | <b>+72.6%</b> | <b>88.3</b> | <b>+81.8%</b> |
| Oracle               | $\{\mathcal{T}\}$                               | 83.7        | 88.2          | 92.9        | 93.9          | 63.3        | 74.1          |

$\mathcal{S}$ : labelled source domain;  $\mathcal{T}$ : labelled target domain;  $\mathcal{T}_C$ : small subset of  $\mathcal{T}$ ;  $\mathcal{T}_P$ : pseudo-labelled target domain

Table 5. Semi-supervised domain adaptation results for Waymo  $\rightarrow$  nuScenes, where 1% of nuScenes data is labelled. The percentages represent the amount of the Direct–Oracle domain gap closed.

**Pseudo-label performance:** Overall, SOAP pseudo-labels improve over the Direct pseudo-labelling baseline by a significant margin. In the nuScenes  $\rightarrow$  Waymo setting (Tab. 2), both architectures receive a 25–30 point improvement in mAP, with over 50% domain gap closed. SOAP is as effective in the Waymo  $\rightarrow$  nuScenes setting (Tab. 3), with both architectures showing a 10-point improvement in mAP and over 30% domain gap closed.

More importantly, SOAP consistently improves object pseudo-labels at different ranges, with the largest improvements observed for objects at the 30–50 m range in all settings, closing 40–60% of the domain gap. We suppose this is because objects farther away from the sensor have sparse point clouds and are sometimes occluded. The results highlight the benefits of full-sequence aggregation, as far objects are densified, and occlusion can be alleviated by aggregating multiple viewpoints. The supplementary material includes qualitative examples that further illustrate the accuracy of SOAP pseudo-labels at long range.

**Adaptation performance:** SOAP complements the SOTA domain adaptation technique ST3D. We observe that when equipped with SOAP pseudo-labels, ST3D + SOAP provides better overall performance than ST3D. In the nuScenes  $\rightarrow$  Waymo setting (Tab. 2), ST3D + SOAP closes 20–30% more domain gap than ST3D. While the difference is smaller in the Waymo  $\rightarrow$  nuScenes setting (Tab. 3), there is still a noticeable improvement, with around 5%

more domain gap closed for mAP and over 10% more domain gap closed for NDS. Moreover, the aforementioned improvement over far objects can also be seen after self-training with ST3D. Using SOAP pseudo-labels achieves significantly higher performance compared to ST3D.

#### 4.4. Semi-supervised domain adaptation

In the semi-supervised domain adaptation setting, where a small amount of target domain annotations are available for training, SOAP can also be used to improve pseudo-label quality. To demonstrate this, we compare SOAP with three methods: Direct, Co-training [21], and SSDA3D [21]. As in the unsupervised case, Direct is where the model is trained only on source domain data. Co-training is where the model is trained with a combination of labelled source and labelled target domain data. SSDA3D is a recent SOTA semi-supervised domain adaptation technique. SSDA3D consists of a pseudo-labelling stage with inter-domain CutMix augmentation to improve pseudo-label quality (which we denote CutMix), followed by a target domain training stage with intra-domain MixUp augmentation as regularization. Additionally, we explore the SSDA3D + SOAP configuration, where we replace the SSDA3D pseudo-labels with SOAP pseudo-labels for second-stage target domain training.

Following SSDA3D [21], we use CenterPoint for experiments in this section and consider 1% sequences la-

|               | 1-frame | 5-frame | SFA | QST | SCP | Overall     |               | Stationary  |               |             |               |             |               |
|---------------|---------|---------|-----|-----|-----|-------------|---------------|-------------|---------------|-------------|---------------|-------------|---------------|
|               |         |         |     |     |     | Level 1     | Level 2       | Level 1     | Level 2       |             |               |             |               |
| (a)           | ✓       |         |     |     |     | 7.0         | -23.5%        | 6.0         | -22.0%        | 5.9         | -21.9%        | 4.9         | -20.5%        |
| (b)           |         | ✓       |     |     |     | 20.4        |               | 17.5        |               | 18.6        |               | 15.6        |               |
| (c)           |         | ✓       | ✓   |     |     | 35.7        | +36.6%        | 31.2        | +35.6%        | 38.0        | +50.1%        | 32.6        | +48.3%        |
| (d)           |         | ✓       | ✓   | ✓   |     | 46.7        | +46.1%        | 41.6        | +46.2%        | 52.1        | +57.7%        | 45.8        | +57.9%        |
| (e)           |         | ✓       | ✓   | ✓   | ✓   | <b>50.9</b> | <b>+53.4%</b> | <b>45.6</b> | <b>+53.8%</b> | <b>57.2</b> | <b>+66.4%</b> | <b>50.6</b> | <b>+67.0%</b> |
| <b>Oracle</b> |         |         |     |     |     | 77.5        |               | 69.7        |               | 76.7        |               | 67.8        |               |

Table 6. Ablation study results for nuScenes → Waymo unsupervised domain adaptation. The percentages represent the amount of domain gap closed relative to the 5-frame baseline detector.

belled in the target domain. Note that, unlike experiments in SSDA3D, we uniformly sample entire sequences instead of individual frames. Following how SSDA3D’s pseudo-labelling model is trained, the SOAP model is trained on both labelled source and labelled target sequences with CutMix, among other standard augmentation, applied to aggregated point clouds. We use the SSDA3D CutMix predictions as sparse predictions to construct final SOAP pseudo-labels.

The main results are shown in Tables 4 and 5.

**Pseudo-label performance:** Compared to the pseudo-labels generated by Co-training and SSDA3D CutMix, SOAP pseudo-labels are much more accurate, closing 85.5% and 69.7% domain gap for nuScenes → Waymo (Tab. 4) and Waymo → nuScenes (Tab. 5), respectively. Similar to the results in the unsupervised settings, the improvement is particularly noticeable for objects at 30–50 m, further illustrating the benefits of full-sequence aggregation.

**Adaptation performance:** SOAP improves the already impressive domain adaptation performance achieved by SSDA3D, closing 5–10% more overall performance gap in both settings. Moreover, we observe the improvements in pseudo-labels for objects at 30–50 m translate to the model after adaptation. In the Waymo → nuScenes setting (Tab. 5), training with SOAP pseudo-labels achieves 10.1% higher mAP and 5.5% higher NDS for 30–50 m objects.

#### 4.5. Ablation study

In this section, we investigate the benefits of QST and SCP in the nuScenes → Waymo unsupervised domain adaptation setting, using the VoxelNeXt architecture. The results are presented in Table 6.

We use the 1-frame and 5-frame detectors as baselines in lines (a) and (b), respectively, and progressively introduce each component of SOAP. In line (c), SFA augments the 5-frame detector with stationary object predictions using aggregated point clouds. The model is trained by naively filtering object speed based on a threshold of 0.2 m/s. Line (d) enables QST, replacing naive filtering. While the SFA pseudo-labels improve over both the 1-frame and 5-frame baselines, especially for stationary objects, it is still sig-

nificantly outperformed by QST. This highlights both the effectiveness of full-sequence aggregation in cross-sensor settings and the importance of constructing robust training labels for stationary objects using QST. Moreover, incorporating SCP in line (e) further improves the AP by over 4%, demonstrating the benefit of exploiting the stationarity of the detected objects.

## 5. Limitations

SOAP has three principal limitations. First, constructing aggregated point clouds requires the point cloud data to be collected sequentially and the ego pose estimates to be available. This is applicable in most current self-driving datasets but may not work in other applications where sequential information is not available. Second, SOAP assumes the ego vehicle—hence the sensor—is moving relative to the static environment. It is not applicable to roadside detection where the sensor stays stationary. Finally, since SOAP is designed to detect stationary objects to augment sparse pseudo-labels, for objects like pedestrians or environments with mostly dynamic objects, SOAP may be less effective. However, in major realistic self-driving datasets [2, 10, 16, 22], we find that at least two thirds of vehicles are stationary at some point in the sequence, making SOAP effective for practical applications. Detailed statistics can be found in the supplementary material.

## 6. Conclusion

We have presented Stationary Object Aggregation Pseudo-labelling (SOAP), a novel method that utilizes full-sequence scene-level aggregation to generate high-quality pseudo-labels for the cross-sensor domain adaptation setting. We have provided extensive evaluation that demonstrates SOAP can provide high-quality pseudo-labels and improves the already impressive results achieved by SOTA methods such as ST3D and SSDA3D.

As future work, we want to explore the benefits of tracking and second-stage refinement, as used by in-domain pseudo-labelling methods, in the domain adaptation setting. It will also be interesting to explore the synergy of SOAP with other domain adaptation approaches.



## References

- [1] Alejandro Barrera, Jorge Beltrán, Carlos Guindel, Jose Antonio Iglesias, and Fernando García. Cycle and semantic consistent adversarial domain adaptation for reducing simulation-to-real domain shift in LiDAR bird’s eye view. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2021. 2
- [2] 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 multi-modal dataset for autonomous driving. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2020. 1, 2, 3, 5, 8
- [3] Xuesong Chen, Shaoshuai Shi, Benjin Zhu, Ka Chun Cheung, Hang Xu, and Hongsheng Li. MPPNet: Multi-frame feature intertwining with proxy points for 3D temporal object detection. In *European Conference on Computer Vision (ECCV)*, pages 680–697. Springer, 2022. 3
- [4] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. VoxelNeXt: Fully sparse voxelnet for 3D object detection and tracking. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, pages 21674–21683, 2023. 2, 6
- [5] Eduardo R. Corral-Soto, Amir Nabatchian, Martin Gerdzhev, and Liu Bingbing. LiDAR few-shot domain adaptation via integrated CycleGAN and 3D object detector with joint learning delay. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 13099–13105, 2021. 2
- [6] Guangyao Ding, Meiyang Zhang, E Li, and Qi Hao. JST: Joint self-training for unsupervised domain adaptation on 2D&3D object detection. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 477–483. IEEE, 2022. 2
- [7] Lue Fan, Yuxue Yang, Yiming Mao, Feng Wang, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Once detected, never lost: Surpassing human performance in offline LiDAR based 3D object detection, 2023. arXiv: 2304.12315. 3
- [8] Jin Fang, Dingfu Zhou, Jingjing Zhao, Chulin Tang, Cheng-Zhong Xu, and Liangjun Zhang. LiDAR-CS dataset: LiDAR point cloud dataset with cross-sensors for 3D object detection, 2023. arXiv: 2301.12515. 2
- [9] Christian Fruhwirth-Reisinger, Michael Opitz, Horst Possegger, and Horst Bischof. FAST3D: Flow-aware self-training for 3D object detectors. In *British Machine Vision Conference (BMVC)*, 2021. 2
- [10] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinisky, W. Jiang, and V. Shet. Level 5 perception dataset 2020. <https://level-5.global/level5/data/>, 2019. 2, 8
- [11] Meelis Kull, Telmo M. Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial Intelligence and Statistics*, pages 623–631. PMLR, 2017. 5
- [12] Zhipeng Luo, Zhongang Cai, Changqing Zhou, Gongjie Zhang, Haiyu Zhao, Shuai Yi, Shijian Lu, Hongsheng Li, Shanghang Zhang, and Ziwei Liu. Unsupervised domain adaptive 3D detection with multi-level consistency. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8866–8875, 2021. 2, 5
- [13] Tao Ma, Xueming Yang, Hongbin Zhou, Xin Li, Botian Shi, Junjie Liu, Yuchen Yang, Zhizheng Liu, Liang He, Yu Qiao, et al. DetZero: Rethinking Offboard 3D object detection with long-term sequential point clouds, 2023. arXiv: 2306.06023. 3
- [14] Charles R. Qi, Yin Zhou, Mahyar Najibi, Pei Sun, Khoa Vo, et al. Offboard 3D object detection from point cloud sequences. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, pages 6134–6144, June 2021. 3
- [15] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, 2021. 3, 5
- [16] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2020. 1, 2, 5, 8
- [17] Darren Tsai, Julie Stephany Berrio, Mao Shan, Eduardo Nebot, and Stewart Worrall. Viewer-centred surface completion for unsupervised domain adaptation in 3D object detection, 2022. arXiv: 2209.06407. 2, 5
- [18] Darren Tsai, Julie Stephany Berrio, Mao Shan, Stewart Worrall, and Eduardo Nebot. See eye to eye: A LiDAR-agnostic 3D detection framework for unsupervised multi-target domain adaptation. *IEEE Robotics and Automation Letters*, 7(3):7904–7911, 2022. 2, 5
- [19] Tianyu Wang, Xiaowei Hu, Zhengzhe Liu, and Chi-Wing Fu. Sparse2Dense: Learning to densify 3D features for 3D object detection, 2022. arXiv: 2211.13067. 2
- [20] Yan Wang, Xiangyu Chen, Yurong You, Li Erran Li, Bharath Hariharan, et al. Train in Germany, test in the USA: Making 3D object detectors generalize. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2020. 2, 5
- [21] Yan Wang, Junbo Yin, Wei Li, Pascal Frossard, R. G. Yang, and Jianbing Shen. SSDA3D: Semi-supervised domain adaptation for 3D object detection from point cloud. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 2, 7
- [22] 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. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 2, 8
- [23] Qiangeng Xu, Yin Zhou, Weiyue Wang, Charles R Qi, and Dragomir Anguelov. SPG: Unsupervised domain adaptation for 3D object detection via semantic point generation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

- [24] Bin Yang, Min Bai, Ming Liang, Wenyuan Zeng, and Raquel Urtasun. Auto4D: Learning to label 4D objects from sequential point clouds, 2021. arXiv: 2101.06586. 3
- [25] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. ST3D++: Denoised self-training for unsupervised domain adaptation on 3D object detection, 2021. arXiv: 2103.05346. 2, 5
- [26] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. ST3D: Self-training for unsupervised domain adaptation on 3D object detection. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2021. 2, 5, 6
- [27] Zetong Yang, Yin Zhou, Zhifeng Chen, and Jiquan Ngiam. 3D-MAN: 3D multi-frame attention network for object detection. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, pages 1863–1872, 2021. 3
- [28] Zeng Yihan, Chunwei Wang, Yunbo Wang, Hang Xu, Chaoqiang Ye, Zhen Yang, and Chao Ma. Learning transferable features for point cloud detection via 3D contrastive co-training. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 21493–21504. Curran Associates, Inc., 2021. 2, 5
- [29] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3D object detection and tracking. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, pages 11784–11793, 2021. 2, 6
- [30] Yurong You, Carlos Andres Diaz-Ruiz, Yan Wang, Wei-Lun Chao, Bharath Hariharan, et al. Exploiting playbacks in unsupervised domain adaptation for 3D object detection in self-driving cars. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2022. 2, 5
- [31] Weichen Zhang, Wen Li, and Dong Xu. SRDAN: Scale-aware and range-aware domain adaptation network for cross-dataset 3D object detection. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2021. 2