IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2023, London, UK. Cite as RA-L paper.

Teachers in concordance for pseudo-labeling of 3D sequential data

Awet Haileslassie Gebrehiwot[1*], Patrik Vacek[1], David Hurych[2],
Karel Zimmermann[1], Patrick Pérez[2], Tomáš Svoboda[1]

***Abstract*—Automatic pseudo-labeling is a powerful tool to tap into large amounts of sequential unlabeled data. It is especially appealing in safety-critical applications of autonomous driving, where performance requirements are extreme, datasets are large, and manual labeling is very challenging. We propose to leverage sequences of point clouds to boost the pseudo-labeling technique in a teacher-student setup via training multiple teachers, each with access to different temporal information. This set of teachers, dubbed *Concordance*, provides higher quality pseudo-labels for student training than standard methods. The output of multiple teachers is combined via a novel pseudo-label confidence-guided criterion. Our experimental evaluation focuses on the 3D point cloud domain and urban driving scenarios. We show the performance of our method applied to 3D semantic segmentation and 3D object detection on three benchmark datasets. Our approach, which uses only 20% manual labels, outperforms some fully supervised methods. A notable performance boost is achieved for classes rarely appearing in training data. Our codes are publicly available on https://github.com/ctu-vras/T-Concord3D.**

I. Introduction

In many machine learning problems, state-of-the-art performance requires supervision via complete annotation of the training data. Therefore, an effective way to increase the performance of a model is to add more annotated training data [13]. However, this approach is neither scalable nor sustainable, requiring thorough manual labeling by human annotators. Annotation tasks such as semantic segmentation of videos and sequences of point clouds are very complex to accomplish. Reducing annotation needs is, therefore, a crucial and active research field. *Pseudo-labeling* [16] has emerged as a versatile and powerful tool. A teacher model trained on a small amount of labeled data is used to annotate lots of unlabeled data automatically. The student model trains on the combination of a small labeled set and a large pseudo-labeled set. This work introduces ways to boost pseudo-labeling when dealing with temporally ordered data streams. The proposed framework is instantiated and evaluated in the context of 3D point-cloud analysis for driving applications. In contrast to unordered data, the temporal ordering of the data grants the student and teacher access to a meaningful temporal context. In particular, the teacher can also access future data in the form of privileged information [33] that is available at the time of pseudo-labeling but not at the students' inference time. Consequently, the teacher can benefit from past and future frames, thus making the most of the temporal consistency over an extended time window. We noticed that the range of this temporal window has a crucial influence on teachers' performance, as it captures different temporal contexts. The complexity of spatio-temporal events in 3D driving scenes would require the teacher to operate simultaneously at different temporal ranges. Learning a large enough teacher capable of modeling the aforementioned complexity would require many labels, contradicting the motivation of easing the annotation. We take a more practical approach where multiple complementary teachers are trained, each operating in its own temporal range with the past and future frames unannotated. These teachers use the *Concordance* to assess the confidence of the extracted pseudo-labels (PLs) and to select the most confident ones for student training eventually.

We experimentally demonstrate that Concordance-based pseudo-labeling (i) achieves competitive performance with state-of-the-art fully supervised methods [36], [20], [37], [39] with only a fraction of labeled data, (ii) outperforms pseudo-labeling methods that do not leverage multiple teachers [13].

Our approach (Fig. 1) only assumes that two sets of sequences are available: The first with the central frames annotated and the second larger and devoid of annotation. Several

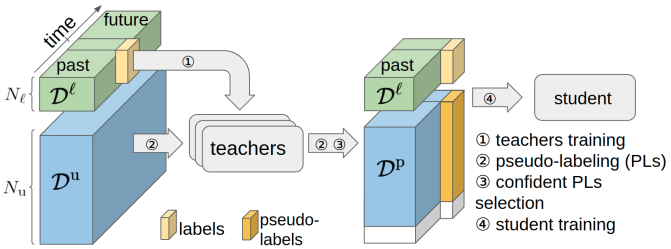


Fig. 1: **Proposed "Concordance of teachers" for pseudo-labeling of sequences**. A set $\mathcal{D}^\ell$ of sequences with a central frame labeled and a larger set $\mathcal{D}^u$ of unannotated ones are available for training; ① Multiple offline teachers are trained with full supervision on $\mathcal{D}^\ell$, each with a different temporal range towards future and past frames; ② The teachers are run on $\mathcal{D}^u$ to produce pseudo-labels (PLs) for central frames; ③ Sequences with the most confident PLs according to Concordance of teachers are selected, forming the pseudo-labeled set $\mathcal{D}^p$. The white box depicts the discarded PLs; ④ The student is trained on $\mathcal{D}^\ell \cup \mathcal{D}^p$, to work online with past and current frames only.

Manuscript received: July, 4, 2022; Revised October, 4, 2022; Accepted November, 17, 2022. This paper was recommended for publication by Editor Vincze, Markus upon evaluation of the Associate Editor and Reviewers' comments.

This work was supported in part by OP VVV MEYS funded project CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics", and by CTU Prague Project SGS22/111/OHK3/2T/13. K. Zimmermann acknowledges CSF Project 20-29531S. The authors want to thank Valeo for its support.

(Corresponding author: A. H. Gebrehiwot)*

[1] *First Author, Second Author, Fourth Author, and Sixth Author are with the Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic*

[2] *Third Author and Fifth Author are with the Valeo.ai*

Digital Object Identifier (DOI): see top of this page.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2023, London, UK. Cite as RA-L paper.

teachers are trained on the first set to predict the label of the middle frame of an input sequence. Once trained, all the offline models run on the second dataset to pseudo-label the central frames. The most promising automatically annotated samples are weighted and added to the first set based on time-aware Concordance sample selection. The resulting large labeled set, with the future frames not available at the input, is used to train the final online model.

We put this framework to work for different spatio-temporal perception tasks on sequences of outdoor point clouds (PCs). We take advantage of the temporal ordering to provide more accurate pseudo-labels than an ordinary PL method would deliver. We demonstrate its superiority on the tasks of 3D detection and 3D semantic segmentation in driving scenes on two architectures [28], [39] and three datasets (see Fig. 2). Our contributions to pseudo-labeling of temporal data are: 1) An effective way to aggregate time-ordered unannotated/annotated 3D scans; Leveraging such privileged information improves teacher's performance for 3D semantic segmentation and object detection tasks. 2) A novel confidence-guided criterion for better pseudo-labels selection and loss function guidance. 3) A novel weighting of pseudo-labels via the Concordance of teachers trained on different temporal ranges.

II. Related Work

Spatial and temporal consistency in point clouds. LiDAR PCs are unstructured data. PointNet architecture [23] directly consumes raw PCs to extract features with permutation invariance and global features for classification. PointNet++ [24] further improves the extraction of local features. The architecture of MeteorNet [15] works with multiple input PCs to extract features from additionally available points with consistent temporal properties. Choy *et al*. [8] use sparse 4D CNN for spatiotemporal perception to improve robustness in detection tasks. Spatial synchronization to the reference scan and adding one additional channel of encoded time lead to further performance gain [11]. Qi *et al*. [25] use temporal information for PC densification based on tracking previously detected objects for automatic data annotation. Conversely to us, they use a top-performance detection model pre-trained in a fully supervised way. We focus on building a set of teacher models (Concordance) from a minimal amount of annotated data and apply distillation through pseudo-labeling.

Knowledge distillation. Training the student model is usually done by distilling knowledge in the feature or output space [9]. Liu *et al*. [18] distill an ensemble of teachers into a single student and extend the idea using different architectures suitable for different tasks as teachers. Cho and Hariharan [7] show that larger models are not necessarily better teachers, mainly due to parameter complexity mismatch. Mirzadeh *et al*. [21] propose a multistep knowledge distillation with an intermediate-sized network to bridge the gap between student and teacher complexity. The benefit of using the teacher network can also come from exploiting privileged information [5]. In our work, we adopt a teacher-student framework and distill future data in sequential frames when training the teacher models.

Semi-supervised learning. Semi-supervised learning approaches have been heavily researched for image recognition, less so for point clouds [13], [38]. Extending an image pseudo-labeling approach [16] to 3D perception [5] has recently shown that automatic labeling of LiDAR data could be leveraged to achieve considerable performance gain. An early approach [30] proposes to extract useful training examples from unlabeled data by exploiting the temporal information in LiDAR scans for classification. Enforcing consistency of model predictions across perturbed versions of unlabelled data [38] proves to be beneficial in 3D object detection. For the task of 3D semantic segmentation, learning with a point-guided contrastive loss [13] increases performance even with fewer ground-truth labels. The authors show that using pseudo-labels and confidence thresholding can help to improve feature learning. This method is compared to ours in Section IV. Our approach focuses on semi-supervised learning using pseudo-labeling and knowledge distillation. It is worth mentioning that a complementary line of work explores approaches such as self-supervised pre-training [35], [22] and domain adaptation [12], [14] to avoid labeling too many samples.

III. Method

Point-cloud notations. A point cloud $X = \{\boldsymbol{x}^k\}_{k=1}^K$ is a finite order-less collection of 3D points, where the number of points K is assumed constant over time to keep the notation simple. We consider symmetric time-ordered PC sequences of the form $X_{-n:n} = (X_{-n}, \dots, X_{-1}, X_0, X_1, \dots, X_n)$ composed of a *reference scan* X_0, preceded by n past scans and followed by n future ones. For the symmetric sequences, the reference scan (frame) is the same as the central one, see Fig. 1. All scans in the sequence are transformed into the coordinate system of the reference one.

Time-aware feature extraction. We build on a modified architectures of the Cylinder3D [39] for semantic segmentation and PointRCNN [28] for object detection. Both architectures consist of MLP modules responsible for attaching rich spatial features to individual scan points, followed by task-dependent modules. In contrast to single-frame perception, we must handle successive scans; therefore, we propose a time-aware extension of their original MLPs.

Given a sequence $X_{-n:n}$ of point clouds, the backbone architecture estimates a feature vector $\boldsymbol{h}(\boldsymbol{x}_0)$ for each point $\boldsymbol{x}_0$ in the reference scan X_0. This feature vector encompasses all contextual information from its spatio-temporal neighborhood $\mathcal{N}(\boldsymbol{x}_0)$ defined as an hourglass-like 3D shape centered in $\boldsymbol{x}_0$:

$$\mathcal{N}(\boldsymbol{x}_0) = \big\{\boldsymbol{x}_t \in X_t : \|\boldsymbol{x}_t - \boldsymbol{x}_0\| \le r(|t|),\ t \in [\![-n, n]\!]\big\}, \quad (1)$$

where r is an increasing function as in [15] (Fig. 3). The feature of each $\boldsymbol{x}_0 \in X_0$ is finally defined as:

$$\boldsymbol{h}(\boldsymbol{x}_0) = \max_{\boldsymbol{x}_t \in \mathcal{N}(\boldsymbol{x}_0)} \big\{\phi(\boldsymbol{x}_t - \boldsymbol{x}_0, t)\big\}, \quad (2)$$

that is, by max-pooling of time-aware pairwise features over the spatio-temporal neighborhood, where ϕ is an MLP with shared weights to be trained, and $t \in [\![-n, n]\!]$.

The construction of the spatio-temporal neighborhoods obeys the intuition that the maximum distance an object can

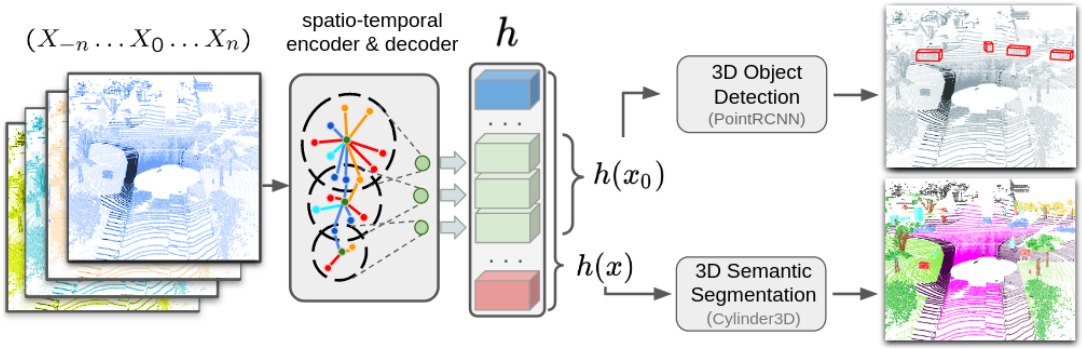


Fig. 2: **Proposed architectures that aggregate sequence of 3D point clouds**. We build a modified version of Cylinder3D for semantic segmentation and PointRCNN for object detection that can aggregate multiple frames inside its spatio-temporal encoder. The output of the spatio-temporal encoder-decoder are extracted point features $\boldsymbol{h}(\boldsymbol{x})$. For object detection, Only features from the reference frame, i.e., $\boldsymbol{h}(\boldsymbol{x_0})$ for $\boldsymbol{x}_0 \in X_0$, are used to train PointRCNN for object detection. Here, we use the full set of features $\boldsymbol{h}(\boldsymbol{x})$ for semantic segmentation due to the setup of the state-of-the-art Cylinder3D architecture.

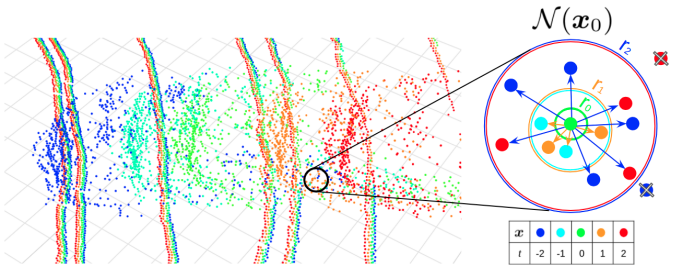


Fig. 3: **Time-aware neighborhood of a point in the reference scan**. Points from different times are presented in different colors, where circled green point $\boldsymbol{x}_0$ is from the (green) reference frame X_0. Its spatio-temporal neighborhood $\mathcal{N}(\boldsymbol{x}_0)$ is composed of all points in scan X_t in a spatial radius $r(|t|)$, for $t = -n, \cdots, n$. Crossed points are excluded from this spatio-temporal neighborhood.

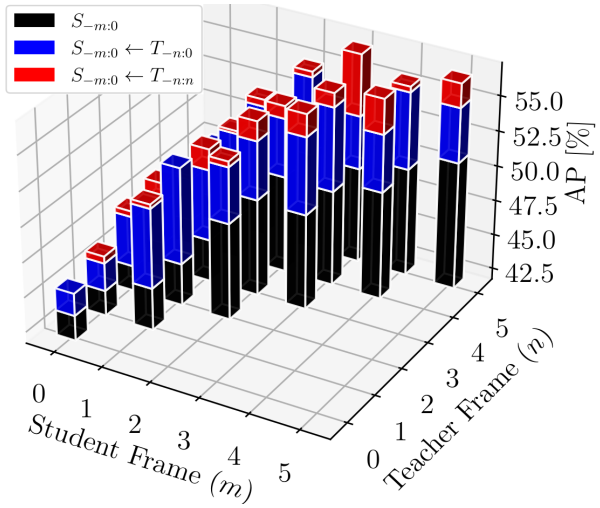


Fig. 4: **Impact of knowledge distillation**. Performance of distilling privileged information from a single teacher into a student in the 3D object detection task.

travel between two scans increases with the object's speed and the temporal separation between the scans. Thus, the maximum spatial distance for grouping points should increase with their temporal distance.

Task-dependent modules. The feature extraction method provides point-wise feature vectors and their corresponding 3D positions for the points in the reference coordinate frame of X_0. The feature vectors are then fed into subsequent task-dependent modules.

3D Object Detection: We consider the 3D detection of vehicles with our multi-frame adaptation of PointRCNN [28]. Here, the bounding box labels are not associated with specific input points but with a specific 3D position. If we consider box labels from all frames, we would not know which points on input are associated with each one. Therefore, we use the bounding box labels only for the reference scan X_0 and mask the box labels from other frames. This differs from semantic segmentation, where each point has its own exclusive class probability in each frame.

3D Semantic Segmentation: We adopt Cylinder3D [39] and extend it into a semi-supervised approach. A trained Cylinder3D classifier semantically labels each point. Here, we use labels for all temporal input instants.

Training data. We consider two types of teacher models for training: (i) $T_{-n:n}$ with access to future frames, and (ii) $T_{-n:0}$, which has access only to past frames.

To understand the effect of distilling a teacher model with access to privileged information $T_{-n:n}$, we have performed an experiment where we train a student model by the pseudo-labels provided by one teacher with and without privileged information. As shown in Fig. 4 distilling a single teacher with privileged information in a student model $S_{-m:0} \leftarrow T_{-n:n}$ provides the best performance (red on top) over the baseline supervised student (black pillars) and over distilling a single teacher without privileged information into a student model $S_{-m:0} \leftarrow T_{-n:0}$ (blue pillars).

Concordance and selection of pseudo-labels. Inspired by our finding from Fig. 4, we want to fully exploit the information contained in the available scan sequences of length $(2n + 1)$.

We propose Concordance of teachers, where we train a set of n teachers with a varying span of the temporal context. It means that teacher $T_{\text{-}1:1}$ is trained on the subsequences $X_{\text{-}1:1}$, teacher $T_{\text{-}2:2}$ on $X_{\text{-}2:2}$ and so on, for the sake of simplicity, we further drop the distinction between these sequences and assume that given a sequence $X \in \mathcal{D}^{\text{u}}$ any network will crop the temporal range appropriately. Given the Concordance of teachers

$$\mathcal{T}^{\{1,\dots,n\}} = \{T_{\text{-}1,1}, T_{\text{-}2,2}, \dots, T_{\text{-}n,n}\}. \tag{3}$$

We independently process every unlabeled sequence $X \in \mathcal{D}^{\text{u}}$ by all trained teachers. Since the nature of outputs is slightly different for the semantic segmentation (point-wise class probabilities) and for the object detection (3D bounding boxes with class probabilities), we split the Concordance description at this point to avoid any misinterpretation.

Concordance for semantic segmentation. We estimate the pseudo-label k^* and its confidence c for each output point. Given the point, each teacher $T \in \mathcal{T}^{\{1,\dots,n\}}$ provides a vector of class probabilities $\hat{\mathbf{y}}^T$. We then estimate the teacher-wise pseudo-labels $k^*(T)$

$$k^*(T) = \underset{k \in K}{\operatorname{argmax}}\, \hat{\mathbf{y}}^T_k. \tag{4}$$

We pay special attention to the teacher with the strongest opinion, the one with the highest output value $\hat{\mathbf{y}}^T_k$. In particular, we denote this teacher as T^*, its pseudo-label k^* and the score of this pseudo-label as $y^* = \hat{\mathbf{y}}^{T^*}_{k^*}$.

We determine the pseudo-label of the given output point as k^*. The confidence of k^* is defined as the weighted sum of two criteria: (i) the score y^* of the pseudo-label and (ii) the number of other teachers T that predict the same class, *i.e.*, for which $k^* = k^*(T)$. The trade-off between these two criteria is determined by a non-negative weight λ as follows:

$$\hat{c} = y^* + \lambda \sum_{T \in \mathcal{T} \setminus T^*} \mathbb{1}[\![k^* = k^*(T)]\!]. \tag{5}$$

To preserve compatibility with training loss Eq. (6), this confidence is clipped: $c = \min(1, \hat{c})$.

Concordance for object detection. Each teacher provides a different set of 3D bounding boxes (bbs) and class probabilities. To calculate final pseudo-label confidence, we need to extract bbs that correspond to a single physical object. We greedily search for clusters of bbs with mutual intersection-over-union above a user-defined threshold. The algorithm starts by building the first cluster from the strongest bb. Once there are no more bbs with a sufficient IoU with the strongest bb, we stop building the cluster, suppress all associated bbs, and continue building a following cluster from the remaining bbs. The class probabilities $\mathbf{y}^T$ corresponding to every single cluster are then used directly to compute the pseudo-label and its confidence by the same procedure as for the semantic segmentation. Following the standard practice in pseudo-labeling (*e.g.*, [19] in object detection or [17] in semantic segmentation), the final selection of pseudo-labels is obtained by thresholding the confidence. Individual pseudo-labels with confidence below the chosen threshold are masked out of the loss function and treated as a *Don't Care* class. The final selection of pseudo-labels and associated data form the training set $\mathcal{D}^{\text{p}}$.

Training the student. The student model only performs inference online and, therefore, is learned only with past input sequences. Following pseudo-labeling through the teacher-student framework [16], [29], we train the student model on both the human-labeled set and the pseudo-labeled one. The loss function for training the student is summed as follows:

$$\mathcal{L} = \frac{1}{M} \sum_{(\mathbf{y},c) \in \mathcal{D}^{\text{p}} \cup \mathcal{D}^{\ell}} c\, \mathcal{L}_{\text{task}}(\hat{\mathbf{y}}, \mathbf{y}), \tag{6}$$

where $\mathbf{y}$ is one hot label encoding vector, $\hat{\mathbf{y}}$ are model predictions, and c is the corresponding confidence from our Concordance selection. Samples collected from the original human-labeled dataset have $c = 1$. The original loss function of the task-dependent module is denoted $\mathcal{L}_{\text{task}}$. All data are sampled from the combined datasets $\mathcal{D}^{\text{p}}$ and $\mathcal{D}^{\ell}$, and M is the number of samples in the union.

IV. Experiments

A. Datasets

We evaluate our approach on the Argoverse dataset [4] for 3D *vehicle* object detection, and SemanticKITTI [1] and nuScenes [3] for 3D semantic segmentation. **Argoverse** provides a large collection of LiDAR sequences with 3D bounding-box labels, from which we utilize the first 10% of human-labeled sequences ($\mathcal{D}^{\ell}$) and 90% being gathered without annotation in $\mathcal{D}^{\text{u}}$ for pseudo-labeling. **SemanticKITTI** provides a large-scale set of driving-scene sequences for 3D semantic segmentation. It consists of 22 sequences that split from 00 to 10 for training (08 reserved for validation) and 11 to 21 for testing. The dataset has two challenges, i.e., *single-scan* with 19 class categories and *multi-scan* with 25 class categories, including 19 from single-scan and six moving-object categories. **nuScenes** contains 1000 scenes with a great diversity of urban traffic and weather conditions. It officially divides the data into 700/150/150 scenes for train/val/test. For our experiment, we cut each sequence into two parts, the first 20% for the human-labeled set $\mathcal{D}^{\ell}$ and the latter 80% for the unlabeled set $\mathcal{D}^{\text{u}}$.

B. 3D Multi-Class Semantic Segmentation

We train a student model on pseudo-labels generated by the concordance of teachers. Here, we utilize Cylinder3D and the output class probabilities for all individual scan points are treated with our confidence-guided criterion (Eqs. 4, 5). Training is done by optimizing the cross-entropy loss and the Lovasz-softmax loss [2] weighted by our confidence-guided criterion, as in Eq. 6. The standard mean Intersection over Union (mIoU) metric is used for evaluation.

Single-scan semantic segmentation. In this experiment, we compare the results of our method with fully-supervised state-of-the-art LiDAR segmentation on SemanticKITTI single-scan test set and nuScenes validation set. Further, we present some qualitative results in Fig. 5a, which show that our model helps to improve the segmentation quality as compared to its

TABLE I: **LiDAR semantic segmentation performance on SemanticKITTI *single-scan* test set.** Our method, $S_{0:0} \leftarrow \mathcal{T}^{\{1,2,3\}}$ ('Ours (20%)') utilizes only 20% of the labeled data, the remaining 80% of training data being automatically annotated, to achieve a performance comparable with state-of-the-art methods. 'Cylinder3D (20%)' denotes Cylinder3D [39] trained, like our method, with only 20% of labeled data; all other results are obtained from the literature, where full (100%) labeled data is used. Performance in IoU percentages, per class and averaged, the higher, the better. Green and red indicate fully-supervised methods that have performance below and above the performance of the proposed method, respectively. '*' means that techniques such as fine-tuning and test-time augmentation (TTA) with flip and rotation are applied.

	mIoU	car	bicycle	motorcycle	truck	o-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	o-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign
RangeNet++ [20]	52.2	91.4	25.7	34.4	25.7	23.0	38.3	38.8	4.8	91.8	65.0	75.2	27.8	87.4	58.6	80.5	55.1	64.6	47.9	55.9
PolarNet [37]	54.3	93.8	40.3	30.1	22.9	28.5	43.2	40.2	5.6	90.8	61.7	74.4	21.7	90.0	61.3	84.0	65.5	67.8	51.8	57.5
SqueezeSegV3 [36]	55.9	92.5	38.7	36.5	29.6	33.0	45.6	46.2	20.1	91.7	63.4	74.8	26.4	89.0	59.4	82.0	58.7	65.4	49.6	58.9
KPConv [31]	58.8	96.0	32.0	42.5	33.4	44.3	61.5	61.6	11.8	88.8	61.3	72.7	31.6	95.0	64.2	84.8	69.2	69.1	56.4	47.4
Cylinder3D [39]	61.8	96.1	54.2	47.6	38.6	45.0	65.1	63.5	13.6	91.2	62.2	75.2	18.7	89.6	61.6	85.4	69.7	69.3	62.6	64.7
(AF)2-S3Net [6]*	69.7	94.5	65.4	86.8	39.2	41.1	80.7	80.4	74.3	91.3	68.8	72.5	53.5	87.9	63.2	70.2	68.5	53.7	61.5	71.0
PVD [10]*	71.2	97.0	67.9	69.3	53.5	60.2	75.1	73.5	50.5	91.8	70.9	77.5	41.0	92.4	69.4	86.5	73.8	71.9	64.9	65.8
Cylinder3D (20%)	51.9	92.1	31.7	28.5	25.1	22.1	49.6	32.4	26.4	86.9	47.2	67.5	12.6	88.7	55.7	83.2	64.4	64.8	53.4	53.9
Ours (20%)	58.9	92.9	46.4	36.6	35.1	27.3	62.4	54.0	24.0	90.0	60.8	72.1	22.2	92.0	65.6	84.6	70.3	63.7	59.3	60.2

TABLE II: **LiDAR semantic segmentation performance on nuScenes valid set**. 'Ours (20%)' utilizes only 20% of the GT annotation, the remaining 80% of training data being automatically annotated, to achieve a performance comparable with state-of-the-art methods. All other results are obtained from the literature, where full (100%) GT annotation is used.

	mIoU	barrier	bicycle	bus	car	construction	motorcycle	pedestrian	trafficcone	trailer	truck	driveable	other	sidewalk	terrain	manmade	vegetation
(AF)2-S3Net [6]	62.2	60.3	12.6	82.3	80.0	20.1	62.0	59.0	49.0	42.2	67.4	94.2	68.0	64.1	68.6	82.9	82.4
RangeNet++ [20]	65.5	66.0	21.3	77.2	80.9	30.2	66.8	69.6	52.1	54.2	72.3	94.1	66.6	63.5	70.1	83.1	79.8
PolarNet [37]	71.0	74.7	28.2	85.3	90.9	35.1	77.5	71.3	58.8	57.4	76.1	96.5	71.1	74.7	74.0	87.3	85.7
PVD [10]	76.0	76.2	40.0	90.2	94.0	50.9	77.4	78.8	64.7	62.0	84.1	96.6	71.4	76.4	76.3	90.3	86.9
CylAsy3D [41]	76.1	76.4	40.3	91.2	93.8	51.3	78.0	78.9	64.9	62.1	84.4	96.8	71.6	76.4	75.4	90.5	87.4
Cylinder3D (20%)	62.0	66.4	13.8	74.7	82.8	16.1	52.1	63.3	48.4	39.3	71.8	95.0	61.6	68.1	71.1	85.0	82.6
Ours (20%)	71.8	73.8	29.3	85.0	90.4	41.6	73.6	74.1	61.1	54.9	78.0	96.1	70.8	73.3	73.9	87.1	85.4

supervised-only counterparts. As shown in Table I, our method $S_{0:0} \leftarrow \mathcal{T}^{\{1,2,3\}}$ trained with only 20% of ground truth (GT) and 80% of pseudo-labeled outperforms all methods based on 3D-to-2D projection with fully-annotated training data [36], [20], [37]. Moreover, our method shows comparable results to voxel partition and 3D convolution-based methods, including fully-supervised Cylinder3D. We made a similar comparison with the fully-supervised state-of-the-art methods on nuScenes validation split. Our method $S_{0:0} \leftarrow \mathcal{T}^{\{1,2,3\}}$ trained with only 20% of GT and 80% of pseudo-labels outperforms some of the fully-supervised models, see Table II.

Multi-scan semantic segmentation. Compared to the single-scan set-up, the multi-scan segmentation in SemanticKITTI has six more categories accounting for moving objects (*car*, *truck*, *other-vehicle*, *person*, *bicyclist* and *motorcyclist*). In this experiment, all methods utilize multiple input point clouds. In Table III, we show that our method, with only 20% of human-labeled training data, outperforms methods that use full manual annotations, namely, DarkNet53 [1] and SqueezesSegv3 [27], and is on par with KPConv [31] and CylAsy3D [41]. Our method outperforms all others on the *moving-car* and *traffic-sign* categories.

Comparison to state-of-the-art semi-supervised method. To further assess the merit of our approach, we compare it to the most recent semi-supervised segmentation work, Guided-Point-SSL [13] on the SemanticKITTI validation set. As shown in Table IV, our method learned from Concordance of teachers, $S_{0:0} \leftarrow \mathcal{T}^{\{1,2,3\}}$, outperforms Guided-Point-SSL with 20%, 30% and 40% labeled data by 1.1, 1.3 and 2.3 mIoU points respectively.

C. 3D Object Detection

The models are trained in the same way as described in PointRCNN [28], except for the confidence-guided criterion and the usage of multiple frames at the input. In Table V, we compare the proposed method to the baseline $S_{\text{-}5:0}$ [28] and to the Mean-Teacher framework (MT) $S_{\text{-}5:0} \leftarrow T_{\text{-}5:0}$ [38]. To reach a fair comparison, we have re-implemented the MT [38] into our architecture and used the classification and regression branches of the original PointRCNN loss function instead of the MT [38] consistency loss. The proposed method $S_{\text{-}5:0} \leftarrow \mathcal{T}^{\{3,4,5\}}$ learned from the concordance of teachers achieves 58.3 AP when trained with 10% human-labeled training data, outperforming the baseline [28] by 7.2 AP and the MT [38]

TABLE III: **LiDAR semantic segmentation performance on SemanticKITTI *multi-scan* test set**. Moving object classes are prefixed with 'mv'; N.B., our model fails to segment 'moving-truck' and 'moving-other' objects as there are no examples of such categories in the 20% labeled split. This is the limitation of the data split.

	mIoU	car	bicycle	motorcycle	truck	o-vehicle	person	road	parking	sidewalk	o-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign	mv-car	mv-truck	mv-other	mv-person	mv-biclist	mv-motor
DarkNet53 [1]	41.6	84.1	30.4	32.9	20.0	20.7	7.5	91.6	64.9	75.3	27.5	85.2	56.5	78.4	50.7	64.8	38.1	53.3	61.5	14.1	15.2	0.2	28.9	37.8
SqueezesSegv3 [27]	43.1	88.5	24.0	26.2	29.2	22.7	6.3	90.1	57.6	73.9	27.1	91.2	66.8	84.0	66.0	65.7	50.8	48.7	53.2	41.2	26.2	36.2	2.3	0.1
KPConv [31]	51.2	93.7	44.9	47.2	42.5	38.6	21.6	86.5	58.4	70.5	26.7	90.8	64.5	84.6	70.3	66.0	57.0	53.9	69.4	0.5	0.5	67.5	67.4	47.2
CylAsy3D [41]	51.5	93.8	67.6	63.3	41.2	37.6	12.9	90.4	66.3	74.9	32.1	92.4	65.8	85.4	72.8	68.1	62.6	61.3	68.1	0.0	0.1	63.1	60.0	0.4
Cylinder3D (20%)	42.1	89.4	35.2	22.9	16.3	15.9	11.6	88.1	53.9	69.2	12.6	88.6	56.8	83.2	65.7	61.3	53.2	59.2	65.8	0.0	0.0	43.3	47.9	12.8
Ours (20%)	47.2	93.0	45.3	35.7	27.4	19.4	14.4	90.5	61.4	75.0	15.6	91.3	62.1	83.3	69.3	64.0	59.7	63.6	77.4	0.0	0.0	64.0	57.5	9.5

TABLE IV: **Semi-supervised learning on SemanticKITTI validation set**. Performance in mIoU (%). 'Guided-Point-SSL' denotes [13] semi-supervised models; '$S_{0:0} \leftarrow \mathcal{T}^{\{1,2,3\}}$' denotes our approach with distillation from the concordance of teachers.

	Labeled data		
method	20%	30%	40%
Guided Point SSL [13]	58.8	59.4	59.9
$S_{0:0} \leftarrow \mathcal{T}^{\{1,2,3\}}$ (Ours)	**59.9**	**60.7**	**62.2**

TABLE V: **Results of 3D object detection**. Detection performance (AP percentage) of proposed students trained with 10% human-labeled training data and of oracle model trained with full (100%) labeled data on Argoverse validation set. '*' indicates reimplementation into our multi-frame PointRCCN architecture.

$S_{\text{-}5:0}$ [28]*	$S_{\text{-}5:0} \leftarrow T_{\text{-}5:0}$ [38]*	$S_{\text{-}5:0} \leftarrow \mathcal{T}^{\{3,4,5\}}$ (Ours)	Oracle
51.1	56.9	58.3	62.6

by 1.4 AP. Moreover, it closes 62.6% of the gap between the baseline [28] and the 'Oracle' (the model $S_{\text{-}5:0}$ trained with 100% GT labeled training data).

D. Implementation Details

We trained models with an ADAM optimizer with a learning rate of 0.001 for 40 epochs on semantic segmentation and 200 and 50 epochs of the RPN and RCNN branches of the object detection model, respectively, using 4 Nvidia A100 GPUs running for 3 days of training for each task. In the object detection task, we subsample each point cloud to 16,384 points from each frame as inputs to the model. We have used three set-abstraction layers ϕ with sizes 4096, 1024, and 128 for our multi-scale time-aware grouping to subsample points into groups. We have used $\lambda = 0.1$ in our experiments.

V. Ablation Studies

Ablation on temporal diversity of teachers. We show that the temporal diversity among teachers overperforms teachers with the same temporal range but with various training initializations. Following the Concordance notation, a set of teachers with the same temporal range is denoted $\mathcal{E}^{\{n,\dots,n\}} = \{T^1_{\text{-}n,n}, \dots, T^K_{\text{-}n,n}\}$, where each randomly initialized teacher $T^k_{\text{-}n,n}$ is operating in *the same* temporal range as others. As shown in Table VI, the student trained by teachers from different temporal ranges outperforms one trained by teachers on the same temporal range, in both 3D object detection and 3D semantic segmentation.

TABLE VI: **Effect of temporal diversity of teachers.** The student trained using concordance of teachers from different temporal ranges outperforms the one trained from the same temporal range but with different initialization in both (a) 3D object detection (Argoverse validation set) and (b) semantic segmentation (SemanticKITTI validation set).

(a) Object detection (AP)

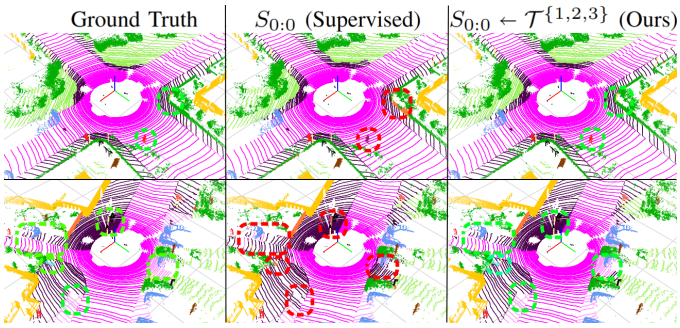
$S_{\text{-}3:0} \leftarrow \mathcal{E}^{\{3,3\}}$	54.3
$S_{\text{-}3:0} \leftarrow \mathcal{T}^{\{2,3\}}$	**55.6**
$S_{\text{-}4:0} \leftarrow \mathcal{E}^{\{4,4\}}$	57.2
$S_{\text{-}4:0} \leftarrow \mathcal{T}^{\{3,4\}}$	**57.7**
$S_{\text{-}5:0} \leftarrow \mathcal{E}^{\{5,5\}}$	58.0
$S_{\text{-}5:0} \leftarrow \mathcal{T}^{\{4,5\}}$	**58.2**

(b) Semantic seg. (mIoU)

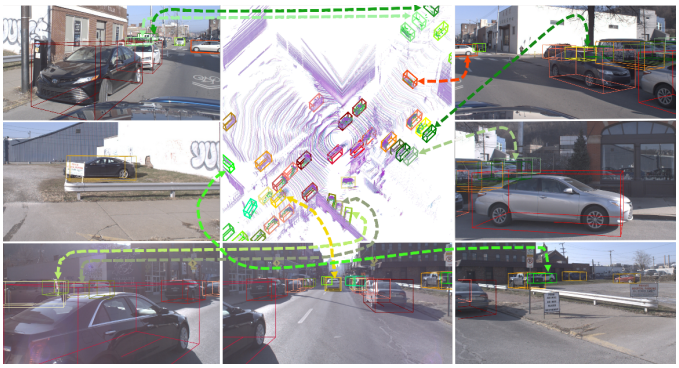
$S_{\text{-}1:0} \leftarrow \mathcal{E}^{\{1,1,1\}}$	59.5
$S_{\text{-}1:0} \leftarrow \mathcal{T}^{\{1,2,3\}}$	**60.6**
$S_{\text{-}2:0} \leftarrow \mathcal{E}^{\{2,2,2\}}$	59.9
$S_{\text{-}2:0} \leftarrow \mathcal{T}^{\{1,2,3\}}$	**60.6**
$S_{\text{-}3:0} \leftarrow \mathcal{E}^{\{3,3,3\}}$	60.0
$S_{\text{-}3:0} \leftarrow \mathcal{T}^{\{1,2,3\}}$	**60.9**

Selection of pseudo-labels. This ablation study demonstrates the benefits of the proposed confidence-guided criterion. The standard baselines here are tuning a single confidence-based threshold (CT) for all pseudo-labels [19], [40], [34]. Models with our proposed selection criterion outperform the CT across multiple confidence thresholds, see Fig. 6.

Effect of labeled and pseudo-labeled dataset ratio. We performed an ablation study to understand the effect of labeled vs. pseudo-labeled training data. We have used the same architecture setup, only varying the amount of labeled and pseudo-labeled data, $|\mathcal{D}^\ell| = 10$, 20, 30, 40, 60, and 100% of training data. As shown in Fig. 7, the gain increases significantly by $\sim$ 10 mIoU when the model uses $|\mathcal{D}^\ell| = 20\%$ of training data compared to $|\mathcal{D}^\ell| = 10\%$. However, the relative gain decreases as the number of labeled data increases to 30 and 40%. This trend shows that the size ratio between $\mathcal{D}^\ell$ and $\mathcal{D}^{\text{p}}$ should be carefully set to achieve adequate performance with the smallest amount of labeled data possible. Moreover, the proposed method, when it uses $|\mathcal{D}^\ell| = 60\%$ of training data (and $|\mathcal{D}^{\text{p}}| = 40\%$), reaches the performance of the fully-



(a) A scene from SemanticKITTI with its GT semantic segmentation (1st column) showing that our method $S_{\text{-}0:0} \leftarrow \mathcal{T}^{\{1,2,3\}}$ (3rd column) can segment person ■, side-walks ■, bicycle ■ and drivable areas ■ better than the supervised baseline based on Cylinder3D (2nd column). Correct and incorrect segmentations are indicated by green and red circles, respectively. Both methods utilize 20% of human-labeled data.



(b) A scene from Argoverse showing that student $S_{\text{-}5:0} \leftarrow \mathcal{T}^{\{3,4,5\}}$ can be robust to severe occlusions. Some of the correctly detected boxes do not contain GT vehicle points since they correspond to vehicles occluded in the current frame and partially occluded in the past frames, as can be observed in the ring of camera images.

Fig. 5: **Qualitative results**. Examples of (a) 3D semantic segmentation and (b) 3D vehicle detection.

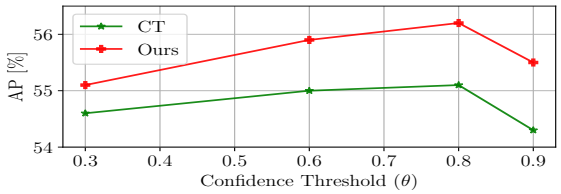


Fig. 6: **Ablation on PL's selection strategy in Argoverse object detection**. Performance as a function of threshold θ. 'Ours' is the proposed confidence-guided criterion and 'CT' the standard confidence-based threshold [19], [40], [34].

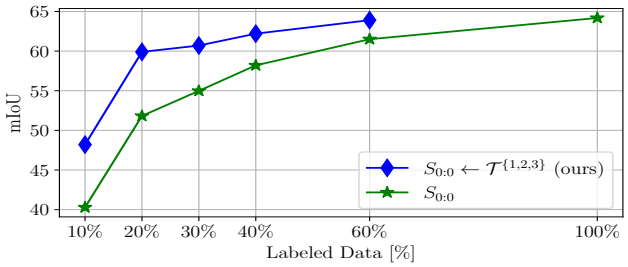


Fig. 7: **Impact of labeling proporiton**. Semantic segmentation performance on SemanticKITTI validation set as a function of labeled data proportion size.

supervised baseline model $S_{0:0}$ which is Cylinder3D [39] trained with 100% of labeled training data; it can be observed by comparing the top-right ending points on the blue and green lines.

Comparision to other teacher-student frameworks. We compare in Table VII our method with other teacher-student approaches such as knowledge distillation (KD) [32], and Ensemble (EN) [26]. We report a comparison using the Cylinder3D $S_{0:0}$ model trained with the methods above using the hyperparameters from Section IV-D. All methods are trained with 20% labeled data. As shown in Table VII, the proposed method significantly outperforms KD [32] and EN [26] baseline teacher-student methods.

TABLE VII: **Comparision to other teacher-student methods**. All methods use $S_{0:0}$, are trained with 20% of labeled data and are evaluated on the SemanticKITTI validation set.

methods	mIoU [%]
Cylinder3D + KD [32]	54.8
Cylinder3D + EN [26]	56.0
Cylinder3D + Ours	59.9

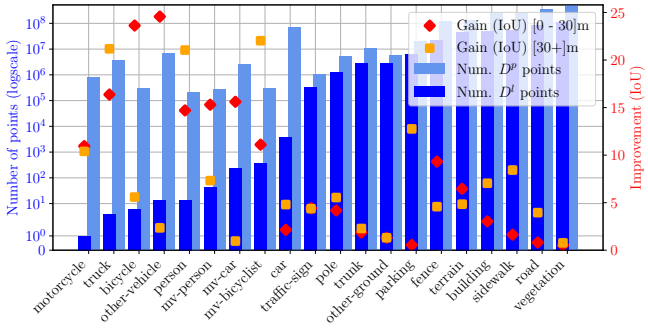


Fig. 8: **Distance-based improvement in class IoUs on SemanticKITTI.** Classes are sorted according to the number of associated labeled training points in $\mathcal{D}^\ell$ and visualized together with pseudo-labeled points ($\mathcal{D}^\text{p}$). We show relative IoU gain within a 30-meter distance from the ego-vehicle ([0-30]m) and at a further distance ([30+]m). Moving objects are prefixed with 'mv'; N.B.: we have omitted classes that do not have points in the labeled split $\mathcal{D}^\ell$.

Distant and close objects. We investigate how multi-scan segmentation is affected by the distance of the points to the ego-vehicle and the number of labeled and pseudo-labeled points for each object class. We compare our method to the baseline [39] on the SemanticKITTI validation set to show the relative gain. Both models are trained with 20% labeled training data. As shown in Fig. 8, a notable performance boost is observed for rarely-appearing classes, especially within 30

meters distance from the ego-vehicle.

Limitations. The SemanticKITTI dataset has a huge data imbalance, and we perform the $\mathcal{D}^{\ell}$ and $\mathcal{D}^{\mathrm{u}}$ split without any relevance to the number of points per object class. We observe cases where no points belong to a specific object category in the labeled set $\mathcal{D}^{\ell}$ resulting in teachers' inability to recognize a particular class, see 'mv-truck' and 'mv-other' in Table III. One should ensure that all the object categories are present in the labeled set $\mathcal{D}^{\ell}$. Secondly, in object detection, since we estimate only objects in the reference frame, we sometimes observe false *false positives*, i.e., detections that are correct but missing in the GT annotation due to no points in the reference frame. The model learns point features from different times and estimate vehicle position in the reference frame. We did not address this issue in evaluating the object detection task.

VI. Conclusion

We propose a novel pseudo-labeling framework that leverages spatio-temporal information from unlabeled sequences of point clouds. We demonstrate its merit in two 3D perception tasks on publicly available datasets. The reported performance gains stem from (i) A better selection of the final pseudo-labels via the concordance of multiple teachers operating at different temporal ranges; (ii) A novel pseudo-label confidence-guided criterion. Thanks to the privileged information available in the different temporal ranges, the Concordance of teachers delivers strong pseudo-labeled samples. Using manual labeling of only 20% of training data, our method achieves state-of-the-art performance in semi-supervised 3D semantic segmentation and competes even with methods that use the full set of labels on this task. By the nature of our pseudo-labeling framework, the proposed approach is complementary to other techniques that use sequential data, and can thus be combined with them to further boost the performance.

References

[1] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In *ICCV*, 2019.

[2] M. Berman, A. R. Triki, and M. B. Blaschko. The Lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *CVPR*, 2018.

[3] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

[4] M. Chang, J. W. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, and J. Hays. Argoverse: 3D tracking and forecasting with rich maps. In *CVPR*, 2019.

[5] X. Chen, B. Mersch, L. Nunes, R. Marcuzzi, I. Vizzo, J. Behley, and C. Stachniss. Automatic labeling to generate training data for online lidar-based moving object segmentation. *IEEE RA-L*, page 6107–6114, 2022.

[6] R. Cheng, R. Razani, E. Taghavi, E. Li, and B. Liu. (af)2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. In *CVPR*, 2021.

[7] J. H. Cho and B. Hariharan. On the efficacy of knowledge distillation. In *ICCV*, 2019.

[8] C. B. Choy, J. Gwak, and S. Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019.

[9] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. In *NIPS DL and RL Workshop*, 2015.

[10] Y. Hou, X. Zhu, Y. Ma, C. C. Loy, and Y. Li. Point-to-voxel knowledge distillation for lidar semantic segmentation. In *CVPR*, 2022.

[11] P. Hu, J. Ziglar, D. Held, and D. Ramanan. What you see is what you get: Exploiting visibility for 3d object detection. In *CVPR*, 2020.

[12] M. Jaritz, T. Vu, R. d. Charette, E. Wirbel, and P. Pérez. xMUDA: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *CVPR*, 2020.

[13] L. Jiang, S. Shi, Z. Tian, X. Lai, S. Liu, C. Fu, and J. Jia. Guided point contrastive learning for semi-supervised point cloud semantic segmentation. In *ICCV*, 2021.

[14] P. Jiang and S. Saripalli. Lidarnet: A boundary-aware domain adaptation model for point cloud semantic segmentation. In *ICRA*, 2021.

[15] Xingyu L., M. Yan, and J. Bohg. Meteornet: Deep learning on dynamic 3d point cloud sequences. In *ICCV*, 2019.

[16] D. Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. In *ICML. Workshop Challenges in Representation Learning*, 2013.

[17] Y. Li, J. Chen, X. Xie, K. Ma, and Y. Zheng. Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation. In *MICCAI*, 2020.

[18] I. J. Liu, J. Peng, and A. G. Schwing. Knowledge flow: Improve upon your teachers. In *ICLR*, 2019.

[19] Y. Liu, C. Ma, Z. He, C. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, and P. Vajda. Unbiased teacher for semi-supervised object detection. In *ICLR*, 2021.

[20] A. Milioto, I. Vizzo, J. Behley, and C. Stachniss. RangeNet++: Fast and accurate LiDAR semantic segmentation. *IROS*, 2019.

[21] S. Mirzadeh, M. Farajtabar, A. Li, and H. Ghasemzadeh. Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher. In *AAAI*, 2019.

[22] L. Nunes, R. Marcuzzi, X. Chen, J. Behley, and C. Stachniss. SegContrast: 3D point cloud feature representation learning through self-supervised segment discrimination. *IEEE RAL*, 2022.

[23] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *CVPR*, 2017.

[24] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, 2017.

[25] C. R. Qi, Y. Zhou, M. Najibi, P. Sun, K. Vo, B. Deng, and D. Anguelov. Offboard 3d object detection from point cloud sequences. In *CVPR*, 2021.

[26] L. Rokach. *Ensemble Learning: Pattern Classification Using Ensemble Methods*. Series in machine perception and artificial intelligence. 2019.

[27] H. Shi, G. Lin, H. Wang, T. Hung, and Z. Wang. SpSequenceNet: Semantic segmentation network on 4D point clouds. In *CVPR*, 2020.

[28] S. Shi, X. Wang, and H. Li. PointRCNN: 3D object proposal generation and detection from point cloud. In *CVPR*, 2019.

[29] W. Shi, Y. Gong, C. Ding, Zhiheng M. Tao, and N. Zheng. Transductive semi-supervised deep learning using min-max features. In *ECCV*, 2018.

[30] A. Teichman and S. Thrun. Tracking-based semi-supervised learning. *IJRR*, 31(7):804–818, 2012.

[31] H. Thomas, C. R. Qi, J. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *ICCV*, 2019.

[32] F. Tung and G. Mori. Similarity-preserving knowledge distillation. In *ICCV*, 2019.

[33] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557, 2009.

[34] H. Wang, Y. Cong, O. Litany, Y. Gao, and L. J. Guibas. 3dioumatch: Leveraging IoU prediction for semi-supervised 3d object detection. In *CVPR*, 2021.

[35] S. Xie, J. Gu, D. Guo, C. R. Qi, L. Guibas, and O. Litany. Pointcontrast: Unsupervised pre-training for 3D point cloud understanding. In *ECCV*, 2020.

[36] C. Xu, B. Wu, Z. Wang, W. Zhan, P. Vajda, K. K., and M. Tomizuka. SqueezeSegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. In *ECCV*, 2020.

[37] Y. Zhang, Z. Zhou, P. David, X. Yue, Z. Xi, B. Gong, and H. Foroosh. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *CVPR*, 2020.

[38] N. Zhao, T. Chua, and G. H. Lee. Sess: Self-ensembling semi-supervised 3d object detection. In *CVPR*, 2020.

[39] H. Zhou, X. Zhu, X. Song, Y. Ma, Z. Wang, H. Li, and D. Lin. Cylinder3d: An effective 3d framework for driving-scene lidar semantic segmentation. *arXiv preprint arXiv:2008.01550*, 2020.

[40] Q. Zhou, C. Yu, Z. Wang, Q. Qian, and H. Li. Instant-teaching: An end-to-end semi-supervised object detection framework. In *CVPR*, 2021.

[41] X. Zhu, H. Zhou, T. Wang, F. Hong, Y. Ma, W. Li, H. Li, and D. Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *CVPR*, 2021.