# LEROjD: Lidar Extended Radar-Only Object Detection

Patrick Palmer[1], Martin Krüger[1], Stefan Schütte[1], Richard Altendorfer[2], Ganesh Adam[2], and Torsten Bertram[1]

[1] Institute of Control Theory and Systems Engineering, TU Dortmund University
{patrick.palmer, martin2.krueger, stefan.schuette, torsten.bertram}@tu-dortmund.de
[2] ZF Group, {richard.altendorfer, ganesh.adam}@zf.com

**Abstract.** Accurate 3D object detection is vital for automated driving. While lidar sensors are well suited for this task, they are expensive and have limitations in adverse weather conditions. 3+1D imaging radar sensors offer a cost-effective, robust alternative but face challenges due to their low resolution and high measurement noise. Existing 3+1D imaging radar datasets include radar and lidar data, enabling cross-modal model improvements. Although lidar should not be used during inference, it can aid the training of radar-only object detectors. We explore two strategies to transfer knowledge from the lidar to the radar domain and radar-only object detectors: 1. multi-stage training with sequential lidar point cloud thin-out, and 2. cross-modal knowledge distillation. In the multi-stage process, three thin-out methods are examined. Our results show significant performance gains of up to 4.2 percentage points in mean Average Precision with multi-stage training and up to 3.9 percentage points with knowledge distillation by initializing the student with the teacher's weights. The main benefit of these approaches is their applicability to other 3D object detection networks without altering their architecture, as we show by analyzing it on two different object detectors. Our code is available at `https://github.com/rst-tu-dortmund/lerojd`.

**Keywords:** 3D Object Detection · 3+1D Imaging Radar · Cross-Modal Object Detection

## 1 Introduction

Environment perception is the first module in each automated driving stack. Multiple sensor modalities, like cameras, lidars, and radars are utilized for this task. Radar sensors are of unique interest in perception due to their robustness against poor lighting, challenging weather conditions like rain or snow, and cost-effectiveness [3, 36, 47]. One exclusive advantage is the ability to measure the relative radial velocity of reflections directly due to the Doppler effect.

While a precise localization of objects is possible with traditional radar sensors without elevation angle measurements [9], it is inherently limited to the horizontal plane. Additionally, it is hard to predict the objects' extent due to

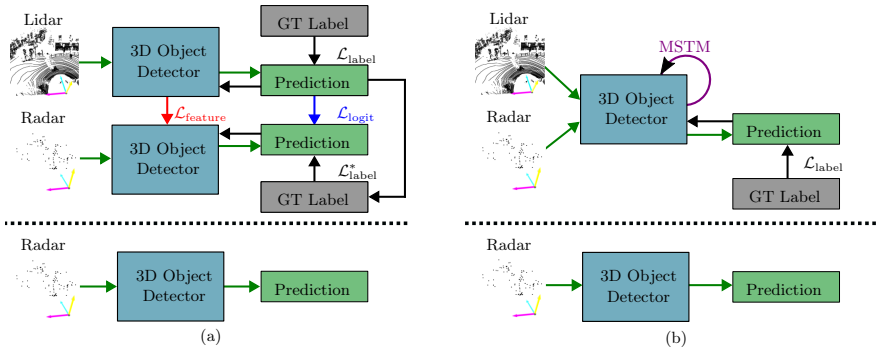



**Fig. 1:** Architecture overview of (a) a knowledge distillation-based method and (b) a multi-stage training method (MSTM) for utilizing lidar data in the training of radar-only object detectors. The ground truth (GT) label is the same for both methods. The diagrams above the dotted line represent the training process, while the diagrams below the dotted line represent inference.

the reflections' sparsity. The introduction of 3+1D high-resolution imaging radar sensors has recently mitigated these limitations, at least partially. In addition to measuring the elevation angle of reflections, the density of measurements is increased [14, 20]. Therefore, approaches that only utilize radar sensors for perception of the environment are of particular interest.

Despite improvements in radar-based object detection, the performance still lags behind other sensor modalities like lidar [36]. One persistent major limitation of 2+1D classic radar and 3+1D imaging radar sensors is the relative sparsity of the point cloud, which limits the detection performance.

Lidar sensors, on the other hand, are well suited for object detection and are therefore frequently employed as a reference for evaluating the performance of different sensor modalities due to their ability to produce an accurate and dense understanding of the scene. Their effectiveness is particularly pronounced in detecting nearby traffic participants without occlusion [53].

All currently available datasets that contain 3+1D imaging radar data additionally accommodate lidar sensor data [4, 8, 32, 35, 36, 40, 63, 65, 69]. The lidar sensor data of these datasets is currently either utilized for labeling, combining multiple sensor modalities for accurate object detection, or comparing the performance of radar-only techniques to another sensor modality. While the majority of series production vehicles may not include lidar sensors due to cost and vehicle design constraints, they will still be available in the training process of learning-based methods. Extending radar-only methods with lidar sensor data in training has been shown to be a viable method for estimating point flow on the imaging radar point cloud [12]. These observations lead to the following research question: *Can lidar sensor data be utilized in the training process of imaging radar-based 3D object detectors to improve the object detection performance on radar-only data during inference?*

To use different sensor modalities in training, transfer learning and knowledge distillation (KD) principles can be utilized. While KD is commonly employed across sensor modalities such as camera images and lidar point clouds, its application between 3+1D imaging radar and lidar sensors remains unexplored. Since lidar and imaging radar share a structurally similar data representation as point clouds, an identical base network with different input modalities can be utilized to transfer knowledge between different input modalities. We investigate two approaches to transfer knowledge from lidar-based to radar-based object detectors: a KD-based and a multi-stage training approach. The main principles of the two methods are visualized in Figure 1.

The contributions of this work are summarized as follows:

- We investigate a combination of lidar and radar sensors in the training stage of object detectors to improve radar-only object detection at inference.
- We investigate three thin-out strategies for lidar point clouds to transfer knowledge from dense lidar to sparse lidar and radar-only object detectors.
- We propose a multi-stage training procedure to transfer knowledge from dense lidar to sparse lidar and, finally, to radar-only object detection.
- We modify and analyze several knowledge distillation-based approaches to transfer knowledge from lidar to radar-only object detectors.

## 2 Related Work

Imaging radar sensors commonly utilize the point cloud as a data representation format instead of the radar tensor, due to its higher computational efficiency. This format is similar to the one used for lidars, enabling the application of object detection methods developed for lidars to radars. Object detection on lidars can be split into two main categories. Point-based methods, like PointNet++ [39], downsample the original point cloud, encode it using a backbone, and finally apply a detection head. Voxel-based methods discretize the point cloud into a 3D grid and apply 3D convolutions to the grid before finally applying a detection head [6, 10, 71]. The main drawback of voxel-based methods is the high memory consumption and the loss of spatial information due to the discretization of the point cloud. To overcome the limitation of high memory consumption, [23, 27] have proposed the PointPillars network. Pillars are a specific form of voxels that span over the full height of the scene and represent the point cloud in a 2D grid.

Methods from the lidar domain have been shown to perform reasonably well on radar data [36, 37]. However, they are limited due to the sparsity of the radar point cloud. SMURF [30] considers two representations of the radar point cloud to address sparsity. Using kernel density estimation, it utilizes pillarization and density features derived from a multi-dimensional Gaussian mixture distribution. RPFA-Net [57] is a PointPillars [23] based network, which introduces a self-attention mechanism to extract global context information from the radar point cloud. RadarMFNet [48] utilizes a multi-frame radar point cloud representation to address the sparsity of the radar point cloud in conjunction with an anchor-based detector and temporal pooling layers.

One way of improving 3D object detection on radar data is the fusion with additional sensor modalities like camera [56, 68, 70], lidar [50] or camera and lidar [5, 13, 55] at the cost of introducing additional sensors at inference.

The concept of knowledge distillation was first introduced by [18]. It consists of two networks, which are labeled as teacher and student. The teacher is a large and complex model, while the student is less complex and more computationally efficient. The student is trained to mimic the teacher network's performance by utilizing the teacher's predictions and the ground truth labels. This has been utilized in the context of 3D object detection by [29, 59, 64] to construct computation time-efficient models that have similar performance as large models. [26] and [22] have extended upon this concept by utilizing KD to extract knowledge from a teacher trained on lidar data to a student who utilizes camera images.

To the best of our knowledge, few studies have investigated the effect of KD considering radar data. A transfer of knowledge from an image-based teacher network to a radar-based student has been shown to improve the performance for the task of people counting [44]. For the task of 3D object detection, [22] has shown that a transfer learning-based approach from a lidar and image-based network to the radar domain results in improved object detection performance for classical 2+1D radar sensors. The main drawback of [22] is that in addition to 3D object labels, instance segmentation labels for the image domain are required. Additionally, [22] introduces multiple sub-networks to derive the KD losses, which makes the model more complex, while we aim for a simpler approach. HiddenGems [12] utilizes lidar point clouds to derive point flow information and train a network to predict the point flow on radar point clouds. See Beyond Seeing [11] utilizes lidar point cloud for feature hallucination on radar point clouds. Both methods require the lidar point cloud only in the training process but require major modifications to the network architecture.

Sampling of point clouds is a broadly explored topic to reduce the complexity of computing large point clouds. Most commonly, farthest point sampling (FPS) [39,42,60], voxelized FPS [41] or random sampling [71] are utilized in the context of 3D object detection. Random sampling has been shown to be beneficial when used for semantic segmentation of point clouds [19] and to outperform farthest point sampling at this task [25]. One limitation of random sampling is that while points at a close distance are kept, points at a far distance are more likely to be removed. Therefore, [7] proposes a polar cylinder balanced random sampling to keep a more balanced distribution of points across the range.

## 3 Method

Two methods of transferring knowledge from lidar-based to radar-based object detectors are investigated: Knowledge Distillation (KD) (Section 3.1), which is modified for the task of transfer learning and our proposed multi-stage training procedure with sequential point cloud thin-out (Section 3.2). Additionally, the utilized thin-out strategies for lidar point cloud (Section 3.3) are described.

### 3.1 Knowledge Distillation

KD is commonly used for two tasks. First, designing computationally efficient models by transferring knowledge from a larger teacher network to a smaller student network [18, 49, 51]. Second, to transfer knowledge across sensor modalities by utilizing different model architectures for the teacher and student [1, 16, 67]. Lidar and radar-based point cloud object detection can utilize the same model structure but with different input modalities. This enables the utilization of KD methods first described for designing computationally efficient models for cross-modal knowledge transfer from lidar- to radar-based object detectors. In this case, the teacher is trained on the full lidar point cloud, while the student is trained on the radar point cloud. Three different loss terms, as described by [59], are employed:

**Logit-KD** is the first, classic type of distilling knowledge described by [18]. For 3D object detection, the logit-KD loss $\mathcal{L}_{\text{logit}}$ is split into a classification $\mathcal{L}_{\text{l-cls}}$ and regression loss $\mathcal{L}_{\text{l-reg}}$. These losses are calculated by comparing the student's and teacher's predictions utilizing the 3d detectors' regression loss and bi-linear interpolation between student and teacher output classes.

**Feature-KD** is widely utilized in 2D object detection [28, 51]. It utilizes a loss term that forces the student network to mimic the teacher's intermediate feature map (feat). A feature mimicking the last layer of the bird's eye view feature encoder, similar to [59], is utilized in this work.

**Label-KD** is a recent distillation approach that leans on the concept of the Logit KD but simplifies and generalizes it. It is first described by [33]. The teacher predictions are filtered by their scores using a score threshold, and an adapted ground truth set is constructed by combining the filtered predictions and the ground truth set. This adapted set is utilized in student training. The loss is split into a classification $\mathcal{L}_{\text{cls}}$ and regression loss $\mathcal{L}_{\text{reg}}$. It replaces $\mathcal{L}_{\text{label}}$ usually calculated on the ground truth set.

The three KD loss terms are combined into a joint loss weighted with $\lambda$

$$\mathcal{L}_{\text{joint}} = \underbrace{\lambda_{\text{l-reg}}\mathcal{L}_{\text{l-reg}} + \lambda_{\text{l-cls}}\mathcal{L}_{\text{l-cls.}}}_{\mathcal{L}_{\text{logit}}} + \underbrace{\lambda_{\text{feat}}\mathcal{L}_{\text{feat}}}_{\mathcal{L}_{\text{feature}}} + \underbrace{\lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}}_{\mathcal{L}^{*}_{\text{label}}} \tag{1}$$

### 3.2 Multi-Stage Lidar Thin-Out Training Procedure

Utilizing a different input data modality for either pre-training a network on a large dataset or utilizing simulated data in the training process of a point cloud-based network has been shown to improve the object detection performance [54]. The multi-stage training method (MSTM) proposed in this work extends upon this by utilizing a Curriculum learning [2] based training procedure by which the network is trained on iteratively sparsified lidar point clouds, similar to [52], and fine-tuned on the radar point cloud. Figure 2 visualizes our multi-stage training procedure. The network is first trained on the full lidar point cloud until convergence. In the following steps, the lidar point cloud is thinned-out iteratively by a factor of 2 and utilized for training a network whose weights are

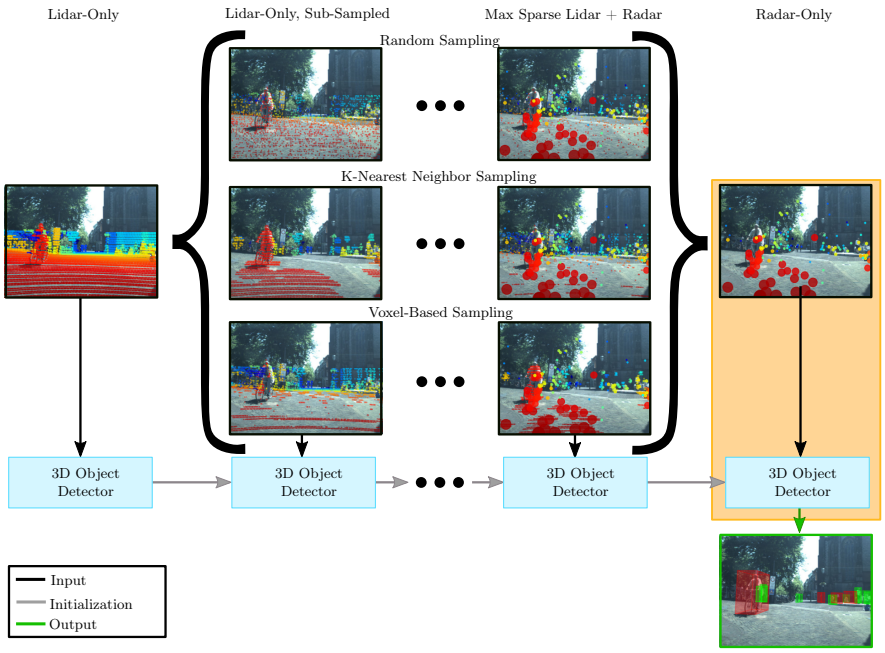



**Fig. 2:** MSTM pipeline. A 3D object detection network is iteratively trained on increasingly sparse lidar data. Three different thin-out strategies are utilized for sparsification. The lidar point cloud is mixed with the radar point cloud in the second to last step. In the last step, the network is only trained on radar data. At inference, only the orange-shaded part is executed. Small points represent lidar points, large points radar points. The color of points corresponds to the distance from the ego-vehicle. The camera image is not used as an input but only aids the visualization.

initialized using the previously trained model. This forces the network to learn features for a good object detection performance on increasingly sparser point clouds. In the second to last step, the lidar point cloud is mixed with the radar points so that the network can translate from the lidar to the radar domain. In the last step, the model is trained only on radar points. A training without multiple stages is called single-stage training method (SSTM) in this work.

In addition to the training where only lidar points are utilized in the first training stages, we investigate the utilization of the radar point cloud in conjunction with the lidar point cloud in all stages. The thin-out of the lidar points remains the same and is mixed with the radar point cloud in each step. This conditions the model on radar from the first step in order to prioritize features in the lidar point cloud that relate to a good object detection on radar-only data.

For mixing lidar and radar point clouds, the voxel or pillar feature encoder is modified to prioritize the radar point cloud in the random sampling process as done by [34]. Otherwise, the vastly higher number of lidar points, even when thinned-out, might lead to the complete exclusion of radar points.

### 3.3 Lidar Thin-Out Strategies

Three different methods for sub-sampling the lidar point cloud are investigated. Figure 2 shows examples for each thin-out stage.

**Random sampling** is the simplest method of sparsifying a point cloud. It neglects the structure and inherent limitations of the point cloud representation, especially for objects far away or with a high degree of occlusion. Neglecting the structure can lead to the complete loss of information for objects represented only by few points.

**K-nearest neighbor sampling** approximates the reflection density distribution of the radar point cloud with the lidar point cloud by only keeping lidar reflections close to radar reflections. This algorithm is described in Algorithm 1. Objects not detected by the radar sensor are, therefore, also not represented by the k-nearest neighbor thinned-out lidar point cloud.

**Voxel-based sampling** aims to reduce the number of points in each area of the point cloud while keeping the general distribution of the point cloud. This is motivated by the fact that radar sensors do not suffer as much from loss of resolution with distance as lidar sensors. The approach is described in Algorithm 2, and executed iteratively for a sequence of sparsification steps.

**Algorithm 1:** K-nearest neighbor sampling

**Input :** Lidar point cloud $\boldsymbol{L} \in \mathbb{R}^{N\times 3}$, radar point cloud $\boldsymbol{R} \in \mathbb{R}^{M\times 3}$
**Output:** Sub-sampled lidar point cloud $\boldsymbol{L_s} \in \mathbb{R}^{K\times 3}$

Calculate the euclidean distance $c_i$ between each lidar point $\boldsymbol{l_i} \in \boldsymbol{L}$ and its nearest radar point from $\boldsymbol{R}$;
Select a share of $K$ lidar points with the smallest distances $\boldsymbol{c}$;
**return** Share of Points $\boldsymbol{L_s} \in \mathbb{R}^{K\times 3}$

**Algorithm 2:** Voxel-based sampling

**Input :** Lidar point cloud $\boldsymbol{L} \in \mathbb{R}^{N\times 3}$
**Output:** Sub-sampled lidar point cloud $\boldsymbol{L_s} \in \mathbb{R}^{K\times 3}$

Initialize $\boldsymbol{L_s}$ with $\boldsymbol{L}$;
Voxelize the lidar point cloud into $v$ voxels;
Calculate the number of points in each voxel $\boldsymbol{P_v}$;
Calculate the minimum number of points per voxel $p_{\min}$, so that at least $0.75\ N$ points are in voxels with more than $p_{\min}$ points;
From each voxel with more than $p_{\min}$ points, choose $p_{\min}$ random points to keep and add remaining points to $\boldsymbol{L_p}$;
Randomly select $0.5\ N$ points from $\boldsymbol{L_p}$ and remove them from $\boldsymbol{L_s}$;
**return** Share of Points $\boldsymbol{L_s} \in \mathbb{R}^{K\times 3}$

Another common sparsification method utilized for imitating low-resolution lidar sensors is layer-based sampling [52, 61]. This approach is not investigated because radar sensors do not capture the environment on a layer basis.

# 4 Experimental Evaluation

## 4.1 Experimental Setup

**Dataset:** All experiments are conducted utilizing the *View-of-Delft* (VoD) dataset [36]. It contains synchronized data of multiple sensor modalities. The 64-layer lidar sensor and the imaging radar are utilized in this work. A point cloud accumulated over 5 frames [36], which has been shown to improve object detection performance compared to no accumulation, is used for radar data [36, 38]. We detected a duplication of identical points in the lidar point cloud, which could adversely affect all sampling methods; thus, we eliminated the duplicated points from the point cloud. Although the VoD dataset is among the best currently available datasets for imaging radar-based object detection, it is limited by its size compared to other automotive datasets without imaging-radar, like [31, 45]. Given the absence of publicly available labels and limited online evaluation for the test dataset, we repurpose the validation dataset as a test set. Consequently, we partition the original training set into a new training set (80 %) and a dedicated validation set (20 %) to ensure robust model training.

**Evaluation Metrics:** The primary performance metric utilized to compare the results is the mean average precision (mAP), as used by [15] [37]. Similar to the evaluation of the Waymo data set [46] in [62], we split the results into two distance bins: short-range (SR): 0-30 m and mid-range (MR): 30-50 m. All experiments are conducted utilizing three different random network initializations that are averaged.

**Training:** Most experiments use the PointPillars model [23] as an object detector with the same configuration as utilized by [36]. For imaging radar data, PointPillars has been shown to perform among the best out of multiple state-of-the-art 3D object detection methods while still performing adequately on lidar data [37]. Furthermore PointPillars is a relevant baseline for radar-specific object detection methods [57, 58]. To show that the proposed MSTM and KD apply to various object detectors, the most promising approaches from the evaluations on PointPillars are evaluated on DSVT-P [17], as an example for a transformer-based model. All SSTM trainings are conducted with an early stopping policy for a maximum of 125 epochs. For the MSTM the initial training on the full lidar point cloud is conducted for 125 epochs, while each refinement step is trained for 30 epochs. All trainings utilize the Adam optimizer [21] and an adapted learning rate scheduler that reaches its maximum learning rate earlier and has a faster descent than the scheduler described by [43]. This improves the object detection performance on radar data.

**Notation:** To distinguish between methods, the following notation is used:

$$\mathcal{T}^{\mathcal{TM}}_{\mathcal{LS}/\mathcal{TO}} \rightarrow \mathcal{T}^{\mathcal{KD}}.$$

The notation is split into two parts. The part left of the arrow represents the data set utilized for pre-training, while the right part represents the data and training method utilized for the last (fine-tuning) training stage of the model.

**Table 1:** Examples of training steps and evaluation data corresponding to the notation, empolying voxel-based sub-sampling. Voxel sampling is shortened to v for legibility.

| Notation | Training | Evaluation |
|---|---|---|
| $\mathrm{L}^{\mathrm{MSTM}}_{\text{1-1/8/v}}$ | $\mathrm{L}_{1/\mathrm{v}} \rightarrow \mathrm{L}_{1/2/\mathrm{v}} \rightarrow \mathrm{L}_{1/4/\mathrm{v}} \rightarrow \mathrm{L}_{1/8/\mathrm{v}}$ | $\mathrm{L}_{1/8/\mathrm{v}}$ |
| $\mathrm{RL}^{\mathrm{MSTM}}_{\text{1-1/8/v}}$ | $\mathrm{RL}_{1/\mathrm{v}} \rightarrow \mathrm{RL}_{1/2/\mathrm{v}} \rightarrow \mathrm{RL}_{1/4/\mathrm{v}} \rightarrow \mathrm{RL}_{1/8/\mathrm{v}}$ | $\mathrm{RL}_{1/8/\mathrm{v}}$ |
| $\mathrm{L}^{\mathrm{MSTM}}_{\text{1-1/16/v}} \rightarrow \mathrm{R}$ | $\mathrm{L}_{1/\mathrm{v}} \rightarrow \mathrm{L}_{1/2/\mathrm{v}} \rightarrow \mathrm{L}_{1/4/\mathrm{v}} \rightarrow \mathrm{L}_{1/8/\mathrm{v}} \rightarrow \mathrm{RL}_{1/16/\mathrm{v}} \rightarrow \mathrm{R}$ | R |
| $\mathrm{RL}^{\mathrm{MSTM}}_{\text{1-1/16/v}} \rightarrow \mathrm{R}$ | $\mathrm{RL}_{1/\mathrm{v}} \rightarrow \mathrm{RL}_{1/2/\mathrm{v}} \rightarrow \mathrm{RL}_{1/4/\mathrm{v}} \rightarrow \mathrm{RL}_{1/8/\mathrm{v}} \rightarrow \mathrm{RL}_{1/16/\mathrm{v}} \rightarrow \mathrm{R}$ | R |
| $\mathrm{RL}^{\mathrm{SSTM}}_{\text{1/4/v}} \rightarrow \mathrm{R}$ | $\mathrm{RL}_{1/4/\mathrm{v}} \rightarrow \mathrm{R}$ | R |

This part is omitted if it is the same as the training dataset. The training data $\mathcal{T}$ can either be lidar (L), radar (R), or mixed radar + lidar data (RL). The training method $\mathcal{TM}$ can denote either MSTM or SSTM. The lidar share $\mathcal{LS}$ is the fraction of the original lidar point cloud that is utilized. Radar-only training always uses the full radar point cloud, therefore $\mathcal{LS}$ is omitted. For the MSTM, this is represented by a range of fractions that are iterated in training. The thin-out method $\mathcal{TO}$ can either be random (rand), k-nearest neighbor (knn) or voxel-based (vox). $\mathcal{KD}$ represents the KD method; this can either be label (lab), logit (log), feature (feat), or a combination of those (joint). $\mathcal{KD}$ is omitted if just an initialization and fine-tuning is utilized. Examples for the training corresponding to specific notations are listed in Table 1.

### 4.2 Evalutation of MSTM on Lidar-only and Mixed Radar + Lidar

To evaluate the applicability of our proposed MSTM to the thinned-out lidar point cloud, we evaluate the training without the last two steps involving radar data. Training is conducted for just the lidar point cloud and the mixed radar + lidar point cloud in all stages. The multi-stage trained network is evaluated after each thin-out stage on the thinned-out lidar (or mixed radar + lidar) point cloud. Thin-out stages up to $1/256$ of the original lidar point cloud are considered due to the lidar point cloud containing fewer points than the radar point cloud at $1/256$ of the original lidar point cloud. The MSTM is compared to the SSTM, which is trained only on the thinned-out point cloud. The results are shown in Fig. 3, complete quantitative results are given in the supplementary.

All thin-out strategies result in an increasing degradation of the detection performance. This is most pronounced when considering a random thin-out, which drops approximately linearly with each thin-out step. K-nearest neighbor and voxel-based sampling consistently perform better than random sampling due to keeping a higher point density in areas around objects, which supports object detection. For k-nearest neighbor sampling, the performance only drops by 0.6 percentage points between $\mathrm{L}^{\mathrm{SSTM}}_{\text{1 / knn}}$ and $\mathrm{L}^{\mathrm{SSTM}}_{\text{1/2 / knn}}$ due to mostly ground points being removed in the first thin-out stage. For voxel-based sampling, after a sharp performance drop in the first thin-out stage, only a relatively slight performance drop is observable until the $1/16$ thin-out stage. In the first stages,

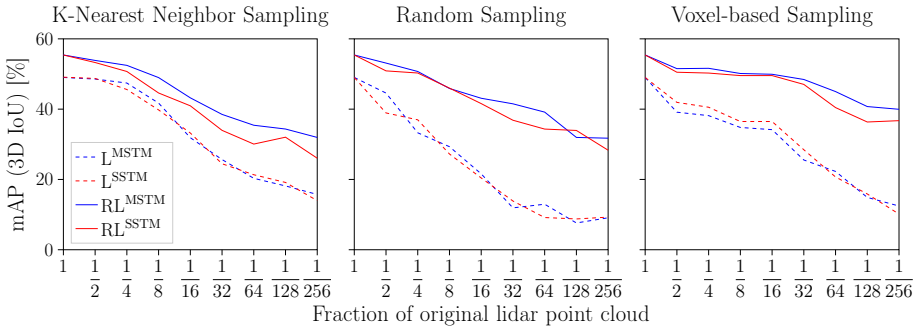



**Fig. 3:** Comparison between SSTM and MSTM for different lidar sampling strategies.

the performance drop can mainly be attributed to the pedestrian class, which drops by 12.8 percentage points in the first thin-out stage. In comparison, the detection performance of the car and cyclist classes only decreased by 0.7 and 3 percentage points, respectively. This can be explained by the voxel size of 1 m in each dimension utilized in the voxel-based sampling. A single pedestrian only occupies a small number of these voxels. When the point cloud gets thinned out, the entire object's structure gets lost, making detection difficult. In contrast, the object's structure and physical appearance can be adequately represented when it occupies more voxels, as observed in the car and cyclist classes. In the final thin-out stage of $\mathrm{L}^{\mathrm{SSTM}}$ and $\mathrm{L}^{\mathrm{MSTM}}$, voxel-based sampling performs worse than k-nearest neighbor sampling because many voxels only consist of ground points or points of surrounding background objects.

The multi-stage training does not result in a useful knowledge transfer and, therefore, significant performance benefit for all considered thin-out strategies.

A contrary behavior is observed on the combined radar + lidar point cloud. The MSTM consistently outperforms the SSTM when using k-nearest neighbor or voxel-based sampling. Knowledge from the dense point cloud can be transferred to the thin radar + lidar point cloud. Additionally, performance is consistently higher than training just on lidar, especially for voxel-based sub-sampling, which performs best at lower thin-out stages. At small thin-out stages, the voxel-based sub-sampling can still represent the whole object space and give meaningful environmental information. At the same time, the radar point cloud is sufficiently dense for object detection.

### 4.3 Evaluation of MSTM with Last Radar-Only Step

This chapter analyses the performance of the MSTM when applied to radar data, as described in Section 3.2. The MSTM is evaluated for two different procedures. Utilizing only the lidar point cloud and utilizing the mixed radar + lidar point cloud in the first stages. The results of MSTM are shown in Table 2. The lidar thin-out stages up to $1/16$ of the original lidar point cloud are considered due to the performance of $\mathrm{L}^{\mathrm{MSTM}}_{\text{1-1/32 / rand}}$ dropping below $\mathrm{R}^{\mathrm{SSTM}}$.

**Table 2:** 3D object detection results on the VoD radar dataset trained using MSTM. All MSTM methods are trained on radar-only data in the last step. The best and second best results are marked in **bold** and underlined, respectively.

| | mAP | | Car | | Pedestrian | | Cyclist | |
|---|---|---|---|---|---|---|---|---|
| Training method | SR | MR | SR | MR | SR | MR | SR | MR |
| $R^{SSTM}$ (baseline) | 36.7 | 11.9 | 45.2 | 18.1 | 17.1 | 7.4 | 47.7 | 10.2 |
| $L^{MSTM}_{1\text{-}1/16/\ rand} \rightarrow R$ | 36.4 | **16.1** | 44.9 | 20.0 | 17.5 | **9.8** | 46.8 | **18.9** |
| $L^{MSTM}_{1\text{-}1/16\ /\ knn} \rightarrow R$ | 35.7 | 14.0 | 45.6 | 23.2 | 16.7 | 7.1 | 44.8 | 11.7 |
| $L^{MSTM}_{1\text{-}1/16/\ vox} \rightarrow R$ | 34.1 | 14.3 | **47.0** | 23.0 | 15.3 | 9.1 | 40.0 | 10.7 |
| $RL^{MSTM}_{1\text{-}1/16/\ rand} \rightarrow R$ | 35.6 | 14.9 | 44.1 | 20.2 | 17.8 | 7.2 | 44.9 | 17.3 |
| $RL^{MSTM}_{1\text{-}1/16\ /\ knn} \rightarrow R$ | 38.2 | 14.7 | 45.5 | **23.9** | **18.8** | 8.2 | 50.3 | 11.9 |
| $RL^{MSTM}_{1\text{-}1/16\ /\ vox} \rightarrow R$ | **39.7** | 15.4 | 45.9 | 22.5 | 18.4 | 9.7 | **54.7** | 13.9 |

When considering the pre-training using just lidar data, it is observable that the overall performance drops in the SR bin. In contrast, the MR bin is improved for all thin-out methods. The best performance is achieved by utilizing the random thin-out strategy. It performs especially well for pedestrians and cyclists in the mid-range, due to the comparably small size of pedestrians and cyclists resulting in a worse radar reflection characteristic. In contrast these objects are detected well by the lidar sensor. Knowledge about the representation of objects can be transferred from the lidar point cloud to the radar point cloud. K-nearest neighbor and voxel-based thin-out do not improve the performance in the SR bin for the pedestrian and cyclist class but perform better in the short and mid-range for the car class which is explained by the large size of cars, resulting in a better representation of cars in the thin point cloud.

When considering the pre-training with the mixed radar + lidar point cloud, a contrary observation is made compared to only using the lidar in pre-training. The overall performance for the random thinned-out lidar point cloud is worse with the mixed point cloud than when only considering the lidar point cloud. One consistent aspect is the good performance of the cyclists in the mid-range, which still surpasses all other methods, excluding $L^{MSTM}_{1\text{-}1/16\ /\ ra}$. However, the k-nearest neighbor and voxel-based thin-out strategies perform better with the mixed point cloud. The voxel-based thin-out strategy achieves the best performance. It performs best on objects in SR, mainly due to its outstanding performance in detecting cyclists, but is limited in the mid-range, getting surpassed by the MSTM with random thin-out. *The MSTM with voxel-based thin-out can increase the object detection performance on the radar point cloud by 3 percentage points in the SR and 3.5 percentage points in the MR.*

### 4.4 Evaluation of Cross-Modality KD

The configuration of the teacher's training data is of particular interest, as the teacher's performance directly influences the student's. The simplest choice is

**Table 3:** 3D object detection results of the teacher with different training sets. Each teacher is evaluated on the test set of the same data configuration as the training set. The best and second best results are marked in **bold** and underlined, respectively.

| | mAP | | Car | | Pedestrian | | Cyclist | |
|---|---|---|---|---|---|---|---|---|
| Data | SR | MR | SR | MR | SR | MR | SR | MR |
| $L_1^{SSTM}$ | 56.5 | 30.0 | 60.7 | 42.4 | 40.1 | 19.5 | 68.5 | 28.1 |
| $RL_1^{SSTM}$ | **61.6** | **45.1** | **61.9** | **46.5** | **44.4** | **34.0** | 78.5 | **54.8** |
| $RL_{1/4\ /\ rand}^{SSTM}$ | 56.9 | 31.1 | 57.2 | 32.4 | 37.2 | 21.3 | 76.4 | 39.5 |
| $RL_{1/4\ /\ knn}^{SSTM}$ | 58.9 | 36.9 | 54.4 | 31.9 | 43.0 | 30.9 | **79.5** | 47.9 |
| $RL_{1/4\ /\ vox}^{SSTM}$ | 56.1 | 39.3 | 60.1 | 41.6 | 32.8 | 28.8 | 75.5 | 47.5 |

to train the teacher solely on the lidar point cloud. Section 4.2 and Section 4.3 show that mixing the radar and lidar point clouds can benefit radar-only object detection. Therefore, teachers trained on a mixed point clouds, in addition to the ones solely trained on lidar, are compared. Mixed point clouds containing 1/4 of the original lidar point cloud are considered, as a closer representation of the teachers to the student's data can lead to a better performance. Thin out of 1/4 is chosen as the radar + lidar detection performance with radar and 1/8 of the lidar points is worse than the SSTM on only lidar data. Table 3 shows the performance of all utilized teachers.

Table 3 shows the results of the teacher network for different training data configurations. It can be observed that, as expected, the mixed radar and lidar point cloud performs the best, only being surpassed by $RL_{1/4\ /\ knn}^{SSTM}$ on the cyclist class in short range. No model can be considered the overall second best among the other training sets. The performance varies between vehicle classes.

The KD is evaluated individually for each KD method and teacher training set. A joint KD is also considered comprised of all three KD losses. All student networks are initialized (Init) with the weights of the teacher network, as it has been shown to improve the student's performance [59]. Additionally, for the transfer learning between datasets, the pre-training utilizing MSTM in Section 4.3 has been shown to improve the performance on the radar dataset.

The results utilizing the KD are shown in Table 4. Just the initialization of the student network with the teacher's weights already results in a performance benefit in the SR with $RL_1^{SSTM}$ and $RL_{1/4\ /\ rand}^{SSTM}$ as teachers and in the MR with all teachers. This is overall only surpassed by the $RL_{1\text{-}1/16\ /\ vox}^{MSTM} \rightarrow R$ showing that the MSTM is substantially better than a simple initialization.

For label-KD, the best-performing models are the ones where the teacher performs the best. Specifically $L_1^{SSTM} \rightarrow R^{lab}$, which is only surpassed by $RL_{1/4\ /\ knn}^{SSTM} \rightarrow R^{lab}$ in the MR. For worse-performing teachers, using label-KD loss does not result in a performance benefit due to it replacing the ground truth label loss. Feature-KD requires a teacher who learns features from radar data. This is observable in the SR performance of models, where the teacher is trained on mixed radar + lidar data. Logit-KD works well on teacher datasets

**Table 4:** 3D object detection results on the VoD radar dataset utilizing the different KDs and teachers trained on different data configurations. Only the mAP over all classes is specified. For each distillation method and range bin, the best and second best results are marked in **bold** and underlined, respectively. The best result for each teacher are marked in cyan for the SR bin and magenta for the MR bin.

| | Init only ($\rightarrow$ R) | | Logit-KD ($\rightarrow$ $R^{log}$) | | Feature-KD ($\rightarrow$ $R^{feat}$) | | Label-KD ($\rightarrow$ $R^{lab}$) | | Joint-KD ($\rightarrow$ $R^{Joint}$) | |
|---|---|---|---|---|---|---|---|---|---|---|
| Teacher data | SR | MR | SR | MR | SR | MR | SR | MR | SR | MR |
| $R_1^{SSTM}$ | 36.7 | 11.9 | - | - | - | - | - | - | - | - |
| $L_1^{SSTM}$ | 34.9 | 13.5 | 32.8 | 12.6 | 34.8 | 13.6 | 36.6 | 13.7 | 36.0 | **13.5** |
| $RL_1^{SSTM}$ | **39.0** | 14.8 | **39.2** | 13.5 | 39.1 | 13.8 | **38.8** | 13.7 | **38.9** | 12.7 |
| $RL_{1/4\ /\ rand}^{SSTM}$ | **39.0** | 14.2 | 38.9 | 11.6 | **39.4** | 11.8 | 34.0 | 12.2 | 36.4 | 12.3 |
| $RL_{1/4\ /\ knn}^{SSTM}$ | 34.8 | 14.6 | 37.2 | **14.7** | 36.6 | 13.0 | 33.8 | 10.9 | 32.9 | 12.9 |
| $RL_{1/4\ /\ vox}^{SSTM}$ | 35.2 | **15.3** | 35.6 | 13.3 | 37.1 | **15.8** | 35.3 | **14.3** | 32.9 | 9.7 |

which closely resemble the radar point cloud and perform well on the teacher set. This results in good performance of $RL_1^{SSTM} \rightarrow R^{log}$. Besides the KD method the thin-out method utilized in the teacher's training influences the student's performance. Random sampling results in good performance in the SR, while voxel-based sampling results in good MR performance. This is somewhat contrary to what has been observed in the MSTM, where random sampling performs the best in the MR and can be explained by the different thin-out stages utilized in the MSTM and KD. Further qualitative results, as well as detailed quantitative results are given in the supplementary. *Overall, initializing the student with the teacher's parameters yields good performance, with further enhancements primarily achievable through feature-KD with a teacher trained on mixed radar + lidar data.*

### 4.5 Evaluation on DSVT as a Transformer-Based Object Detector

Table 5 shows the results of the MSTM and KD on DSVT-P [17] for selected methods. Similar effects, as observed for PointPillars, are seen on DSVT-P. Initializing the student with the teacher's weights and the MSTM led to a performance benefit. However, contrary to PointPillars, the KD approach does not contribute to any improvements for DSVT-P.

### 4.6 Limitations

The best-performing methods in this work apply only to detectors that share an architecture with the target model, as a direct transfer of weights is performed for the best performance. Different models may require different levels and steps in the thinning process. These choices are additional tuning parameters that must be selected appropriately to maximize the benefit of transferring knowledge.

**Table 5:** Radar-only detection performance utilizing DSVT-P as an object detector. The best and second best results are marked in **bold** and underlined, respectively.

| | mAP | | Car | | Pedestrian | | Cyclist | |
|---|---|---|---|---|---|---|---|---|
| Training method | SR | MR | SR | MR | SR | MR | SR | MR |
| $R^{SSTM}$ | 38.3 | 13.0 | 42.5 | 20.2 | 21.9 | **12.3** | 50.5 | 6.6 |
| $RL^{MSTM}_{1\text{-}1/16\ /\ vox} \rightarrow R$ | **41.5** | **15.6** | **47.8** | **23.6** | **24.5** | 11.0 | **52.3** | **12.4** |
| $RL^{SSTM}_{1/4\ /\ vox} \rightarrow R$ | 38.8 | 13.2 | 44.9 | 18.8 | 23.3 | 11.2 | 48.2 | 9.7 |
| $RL^{SSTM}_{1/4\ /\ vox} \rightarrow R^{feat}$ | 34.7 | 11.3 | 44.0 | 8.8 | 20.5 | 2.8 | 39.7 | 12.7 |

## 5 Conclusion

In this paper, we investigated two methods to transfer knowledge from lidar-based object detectorss to radar-only object detection. First, MSTM with sequential sub-sampling of the lidar point cloud, and second, a KD-based approach. For the MSTM, we have investigated three thin-out strategies for the lidar point cloud. These thin-out strategies are also analyzed for the training of the KD teacher network. Both methods can substantially improve the object detection performance of a radar-only object detector. The MSTM with voxel-based thin-out performs the best overall and can improve detection performance by up to 3.5 percentage points. For the KD methods, it is shown that initializing the student with the teacher's parameters, especially a teacher trained on mixed lidar and radar data, can improve the object detection performance on radar-only data, with further enhancement primarily achieved by utilizing feature-KD.

In *future work*, the applicability to further 3D object detection networks and the behavior with more advanced knowledge distillation like the ones utilized by [22] could be investigated. Due to different effects being noticed with the MSTM and the KD methods, combining both methods could be investigated by utilizing the MSTM as a teacher. To overcome the limitations of choosing a strict thin-out strategy, a learnable point cloud thin-out method [24, 66] can be used.

## References

1. Afouras, T., Chung, J.S., Zisserman, A.: Asr is all you need: Cross-modal distillation for lip reading. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2143–2147. IEEE (2020)
2. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning. pp. 41–48 (2009)
3. Brisken, S., Ruf, F., Höhne, F.: Recent evolution of automotive imaging radar and its information content. IET Radar, Sonar & Navigation **12**(10), 1078–1081 (2018)
4. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous

driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
5. Chen, X., Zhang, T., Wang, Y., Wang, Y., Zhao, H.: Futr3d: A unified sensor fusion framework for 3d detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 172–181 (2023)
6. Chen, Y., Liu, J., Zhang, X., Qi, X., Jia, J.: Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21674–21683 (June 2023)
7. Cheng, H., Han, X., Jiang, H., He, D., Xiao, G.: Pcb-randnet: Rethinking random sampling for lidar semantic segmentation in autonomous driving scene. arXiv preprint arXiv:2209.13797 (2022)
8. Choi, M., Yang, S., Han, S., Lee, Y., Lee, M., Choi, K.H., Kim, K.S.: Msc-rad4r: Ros-based automotive dataset with 4d radar. IEEE Robotics and Automation Letters **8**(11), 7194–7201 (2023)
9. Danzer, A., Griebel, T., Bach, M., Dietmayer, K.: 2d car detection in radar data with pointnets. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC). pp. 61–66 (2019)
10. Deng, J., Shi, S., Li, P., Zhou, W., Zhang, Y., Li, H.: Voxel r-cnn: Towards high performance voxel-based 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1201–1209 (2021)
11. Deng, J., Chan, G., Zhong, H., Lu, C.X.: See beyond seeing: Robust 3d object detection from point clouds via cross-modal feature augmentation. arXiv preprint arXiv:2309.17336 (2023)
12. Ding, F., Palffy, A., Gavrila, D.M., Lu, C.X.: Hidden gems: 4d radar scene flow learning using cross-modal supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9340–9349 (June 2023)
13. Drews, F., Feng, D., Faion, F., Rosenbaum, L., Ulrich, M., Gläser, C.: Deepfusion: A robust and modular 3d object detector for lidars, cameras and radars. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 560–567. IEEE (2022)
14. Engels, F., Heidenreich, P., Wintermantel, M., Stäcker, L., Al Kadi, M., Zoubir, A.M.: Automotive radar signal processing: Research directions and practical challenges. IEEE Journal of Selected Topics in Signal Processing **15**(4), 865–878 (2021)
15. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
16. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2827–2836 (2016)
17. Haiyang Wang, Chen Shi, S.S.M.L.S.W.D.H.B.S., Wang, L.: Dsvt: Dynamic sparse voxel transformer with rotated sets. In: CVPR (2023)
18. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. Deep Learning Workshop at NIPS (2014)
19. Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A.: Randla-net: Efficient semantic segmentation of large-scale point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11108–11117 (2020)
20. Jiang, M., Xu, G., Pei, H., Feng, Z., Ma, S., Zhang, H., Hong, W.: 4d high-resolution imagery of point clouds for automotive mmwave radar. IEEE Transactions on Intelligent Transportation Systems pp. 1–15 (2023)

21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. International Conference for Learning Representations (2015)
22. Klingner, M., Borse, S., Kumar, V.R., Rezaei, B., Narayanan, V., Yogamani, S., Porikli, F.: X3kd: Knowledge distillation across modalities, tasks and stages for multi-camera 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13343–13353 (June 2023)
23. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
24. Lang, I., Manor, A., Avidan, S.: Samplenet: Differentiable point cloud sampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7578–7588 (2020)
25. Li, D., Wei, Y., Zhu, R.: A comparative study on point cloud down-sampling strategies for deep learning-based crop organ segmentation. Plant Methods **19**(1), 124 (2023)
26. Li, J., Lu, M., Liu, J., Guo, Y., Du, Y., Du, L., Zhang, S.: Bev-lgkd: A unified lidar-guided knowledge distillation framework for multi-view bev 3d object detection. IEEE Transactions on Intelligent Vehicles pp. 1–10 (2023)
27. Li, J., Luo, C., Yang, X.: Pillarnext: Rethinking network designs for 3d object detection in lidar point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17567–17576 (June 2023)
28. Li, Q., Jin, S., Yan, J.: Mimicking very efficient network for object detection. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 6356–6364 (2017)
29. Li, Z., Li, Y., Wang, Y., Xie, G., Qu, H., Lyu, Z.: A lightweight model for 3d point cloud object detection. Applied Sciences **13**(11) (2023)
30. Liu, J., Zhao, Q., Xiong, W., Huang, T., Han, Q.L., Zhu, B.: Smurf: Spatial multi-representation fusion for 3d object detection with 4d imaging radar. IEEE Transactions on Intelligent Vehicles pp. 1–14 (2023)
31. Mao, J., Minzhe, N., Jiang, C., liang, h., Chen, J., Liang, X., Li, Y., Ye, C., Zhang, W., Li, Z., Yu, J., XU, C., Xu, H.: One million scenes for autonomous driving: Once dataset. In: Vanschoren, J., Yeung, S. (eds.) Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. vol. 1 (2021)
32. Meyer, M., Kuschk, G.: Automotive radar dataset for deep learning based 3d object detection. In: 2019 16th european radar conference (EuRAD). pp. 129–132. IEEE (2019)
33. Nguyen, C.H., Nguyen, T.C., Tang, T.N., Phan, N.L.: Improving object detection by label assignment distillation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1005–1014 (January 2022)
34. Nobis, F., Shafiei, E., Karle, P., Betz, J., Lienkamp, M.: Radar voxel fusion for 3d object detection. Applied Sciences **11**(12) (2021)
35. Paek, D.H., Kong, S.H., Wijaya, K.T.: K-radar: 4d radar object detection for autonomous driving in various weather conditions. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022)
36. Palffy, A., Pool, E., Baratam, S., Kooij, J.F.P., Gavrila, D.M.: Multi-class road user detection with 3+1d radar in the view-of-delft dataset. IEEE Robotics and Automation Letters **7**(2), 4961–4968 (2022)

37. Palmer, P., Krueger, M., Altendorfer, R., Adam, G., Bertram, T.: Reviewing 3d object detectors in the context of high-resolution 3+1d radar. In: Workshop on 3D Vision and Robotics at the Conference on Computer Vision and Pattern Recognition 2023 (2023)
38. Palmer, P., Krueger, M., Altendorfer, R., Bertram, T.: Ego-motion estimation and dynamic motion separation from 3d point clouds for accumulating data and improving 3d object detection. In: AmE 2023 – Automotive meets Electronics; 14. GMM Symposium. pp. 86–91 (2023)
39. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
40. Rebut, J., Ouaknine, A., Malik, W., Pérez, P.: Raw high-definition radar for multi-task learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17021–17030 (June 2022)
41. Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H.: Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10529–10538 (2020)
42. Shi, S., Wang, X., Li, H.: Pointrcnn: 3d object proposal generation and detection from point cloud. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–779 (2019)
43. Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates. In: Artificial intelligence and machine learning for multi-domain operations applications. vol. 11006, pp. 369–386. SPIE (2019)
44. Stephan, M., Hazra, S., Santra, A., Weigel, R., Fischer, G.: People counting solution using an fmcw radar with knowledge distillation from camera data. In: 2021 IEEE Sensors. pp. 1–4 (2021)
45. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
46. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020)
47. Sun, S., Zhang, Y.D.: 4d automotive radar sensing for autonomous vehicles: A sparsity-oriented approach. IEEE Journal of Selected Topics in Signal Processing **15**(4), 879–891 (2021)
48. Tan, B., Ma, Z., Zhu, X., Li, S., Zheng, L., Chen, S., Huang, L., Bai, J.: 3d object detection for multi-frame 4d automotive millimeter-wave radar point cloud. IEEE Sensors Journal (2022)
49. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1365–1374 (2019)
50. Wang, L., Zhang, X., Li, J., Xv, B., Fu, R., Chen, H., Yang, L., Jin, D., Zhao, L.: Multi-modal and multi-scale fusion 3d object detection of 4d radar and lidar for autonomous driving. IEEE Transactions on Vehicular Technology **72**(5), 5628–5641 (2023)
51. Wang, T., Yuan, L., Zhang, X., Feng, J.: Distilling object detectors with fine-grained feature imitation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4933–4942 (2019)

52. Wei, Y., Wei, Z., Rao, Y., Li, J., Zhou, J., Lu, J.: Lidar distillation: Bridging the beam-induced domain gap for 3d object detection. In: European Conference on Computer Vision. pp. 179–195. Springer (2022)
53. Wu, H., Wen, C., Shi, S., Li, X., Wang, C.: Virtual sparse convolution for multimodal 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21653–21662 (June 2023)
54. Xiao, A., Huang, J., Guan, D., Zhan, F., Lu, S.: Transfer learning from synthetic to real lidar point cloud for semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2795–2803 (2022)
55. Xiao, Y., Liu, Y., Luan, K., Cheng, Y., Chen, X., Lu, H.: Deep lidar-radar-visual fusion for object detection in urban environments. Remote Sensing **15**(18) (2023)
56. Xiong, W., Liu, J., Huang, T., Han, Q.L., Xia, Y., Zhu, B.: Lxl: Lidar excluded lean 3d object detection with 4d imaging radar and camera fusion. IEEE Transactions on Intelligent Vehicles pp. 1–14 (2023)
57. Xu, B., Zhang, X., Wang, L., Hu, X., Li, Z., Pan, S., Li, J., Deng, Y.: Rpfa-net: a 4d radar pillar feature attention network for 3d object detection. In: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). pp. 3061–3066 (2021)
58. Yan, Q., Wang, Y.: Mvfan: Multi-view feature assisted network for 4d radar object detection. In: International Conference on Neural Information Processing. pp. 493–511. Springer (2023)
59. Yang, J., Shi, S., Ding, R., Wang, Z., Qi, X.: Towards efficient 3d object detection with knowledge distillation. In: Advances in Neural Information Processing Systems. vol. 35, pp. 21300–21313. Curran Associates, Inc. (2022)
60. Yang, Z., Sun, Y., Liu, S., Jia, J.: 3dssd: Point-based 3d single stage object detector. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11040–11048 (2020)
61. You, Y., Wang, Y., Chao, W.L., Garg, D., Pleiss, G., Hariharan, B., Campbell, M., Weinberger, K.Q.: Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. In: ICLR (2020)
62. Zamanakos, G., Tsochatzidis, L., Amanatiadis, A., Pratikakis, I.: A comprehensive survey of lidar-based 3d object detection methods with deep learning for autonomous driving. Computers & Graphics **99**, 153–181 (2021)
63. Zhang, J., Zhuge, H., Liu, Y., Peng, G., Wu, Z., Zhang, H., Lyu, Q., Li, H., Zhao, C., Kircali, D., Mharolkar, S., Yang, X., Yi, S., Wang, Y., Wang, D.: Ntu4dradlm: 4d radar-centric multi-modal dataset for localization and mapping. In: 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC) (2023)
64. Zhang, L., Dong, R., Tai, H.S., Ma, K.: Pointdistiller: Structured knowledge distillation towards efficient and compact 3d detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21791–21801 (June 2023)
65. Zhang, X., Wang, L., Chen, J., Fang, C., Yang, L., Song, Z., Yang, G., Wang, Y., Zhang, X., Yang, Q., Li, J.: Dual radar: A multi-modal dataset with dual 4d radar for autononous driving. arXiv preprint arXiv:2310.07602 (2023)
66. Zhang, Y., Hu, Q., Xu, G., Ma, Y., Wan, J., Guo, Y.: Not all points are equal: Learning highly efficient point-based detectors for 3d lidar point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18953–18962 (June 2022)
67. Zhao, M., Li, T., Abu Alsheikh, M., Tian, Y., Zhao, H., Torralba, A., Katabi, D.: Through-wall human pose estimation using radio signals. In: Proceedings of

the IEEE conference on computer vision and pattern recognition. pp. 7356–7365 (2018)
68. Zheng, L., Li, S., Tan, B., Yang, L., Chen, S., Huang, L., Bai, J., Zhu, X., Ma, Z.: Rcfusion: Fusing 4-d radar and camera with bird's-eye view features for 3-d object detection. IEEE Transactions on Instrumentation and Measurement **72**, 1–14 (2023)
69. Zheng, L., Ma, Z., Zhu, X., Tan, B., Li, S., Long, K., Sun, W., Chen, S., Zhang, L., Wan, M., Huang, L., Bai, J.: Tj4dradset: A 4d radar dataset for autonomous driving. In: 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC). pp. 493–498 (2022)
70. Zhou, T., Chen, J., Shi, Y., Jiang, K., Yang, M., Yang, D.: Bridging the view disparity between radar and camera features for multi-modal fusion 3d object detection. IEEE Transactions on Intelligent Vehicles **8**(2), 1523–1535 (2023)
71. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4490–4499 (2018)