

Hidden Gems: 4D Radar Scene Flow Learning Using Cross-Modal Supervision

Ding, Fangqiang ; Palffy, Andras; Gavrilă, Dariu; Lu, Chris Xiaoxuan

DOI

[10.1109/CVPR52729.2023.00901](https://doi.org/10.1109/CVPR52729.2023.00901)

Publication date

2023

Document Version

Final published version

Published in

Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Citation (APA)

Ding, F., Palffy, A., Gavrilă, D., & Lu, C. X. (2023). Hidden Gems: 4D Radar Scene Flow Learning Using Cross-Modal Supervision. In L. O'Conner (Ed.), *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 9340-9349). IEEE.
<https://doi.org/10.1109/CVPR52729.2023.00901>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' - Taverne project

<https://www.openaccess.nl/en/you-share-we-take-care>

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.

Hidden Gems: 4D Radar Scene Flow Learning Using Cross-Modal Supervision

Fangqiang Ding¹ Andras Palffy²
¹University of Edinburgh
 {fding, xiaoxuan.lu}@ed.ac.uk

Dariu M. Gavrilă² Chris Xiaoxuan Lu^{1,*}
²Delft University of Technology
 {a.palffy, d.m.gavrila}@tudelft.nl

Abstract

This work proposes a novel approach to 4D radar-based scene flow estimation via cross-modal learning. Our approach is motivated by the co-located sensing redundancy in modern autonomous vehicles. Such redundancy implicitly provides various forms of supervision cues to the radar scene flow estimation. Specifically, we introduce a multi-task model architecture for the identified cross-modal learning problem and propose loss functions to opportunistically engage scene flow estimation using multiple cross-modal constraints for effective model training. Extensive experiments show the state-of-the-art performance of our method and demonstrate the effectiveness of cross-modal supervised learning to infer more accurate 4D radar scene flow. We also show its usefulness to two subtasks - motion segmentation and ego-motion estimation. Our source code will be available on <https://github.com/Toytiny/CMFlow>.

1. Introduction

Scene flow estimation is to obtain a 3D motion vector field of the static and dynamic environment relative to an ego-agent. In the context of self-driving, scene flow is a key enabler to navigational safety in dynamic environments by providing holistic motion cues to multiple subtasks, such as ego-motion estimation, motion segmentation, point cloud accumulation, multi-object tracking, etc.

Driven by the recent successes of deep neural networks in point cloud processing [22, 35, 36, 40, 51, 54], predominant approaches to scene flow estimation from point clouds adopt either fully- [12, 27, 34, 49, 52] or weakly- [10, 11] supervised learning, or only rely on self-supervised signals [2, 9, 21, 24, 31]. For supervised ones, the acquisition of scene flow annotations is costly and requires tedious and intensive human labour. In contrast, self-supervised learning methods require no annotations and can exploit the inherent spatio-temporal relationship and constraints in the input data to bootstrap scene flow learning. Neverthe-

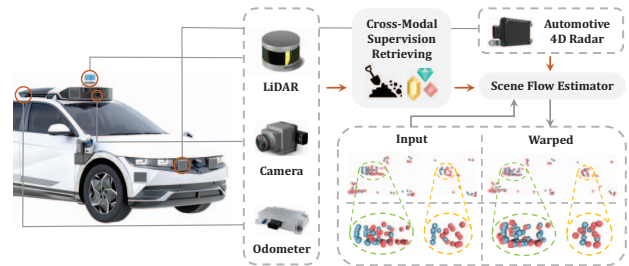


Figure 1. Cross-modal supervision cues are retrieved from co-located odometer, LiDAR and camera sensors to benefit 4D radar scene flow learning. The source point cloud (red) is warped with our estimated scene flow and gets closer to the target one (blue).

less, due to the implicit supervision signals, self-supervised learning performance is often secondary to the supervised ones [10, 23, 49], and they fail to provide sufficiently reliable results for safety-critical autonomous driving scenarios.

These challenges become more prominent when it comes to 4D radar scene flow learning. 4D automotive radars receive increasing attention recently due to their robustness against adverse weather and poor lighting conditions (vs. camera), availability of object velocity measurements (vs. camera and LiDAR) and relatively low cost (vs. LiDAR) [5, 30, 32, 41]. However, 4D radar point clouds are significantly sparser than LiDARs' and suffer from non-negligible multi-path noise. Such low-fidelity data significantly complicates the point-level scene flow annotation for supervised learning and makes it difficult to rely exclusively on self-supervised training-based methods [9] for performance and safety reasons.

To find a more effective framework for 4D radar scene flow learning, this work aims to exploit cross-modal supervision signals in autonomous vehicles. Our motivation is based on the fact that autonomous vehicles today are equipped with multiple heterogeneous sensors, e.g., LiDARs, cameras and GPS/INS, which can provide complementary sensing and redundant perception results for each other, jointly safeguarding the vehicle operation in complex urban traffic. This *co-located perception redundancy*

*Corresponding author: Chris Xiaoxuan Lu (xiaoxuan.lu@ed.ac.uk)

can be leveraged to provision multiple supervision cues that bootstrap radar scene flow learning. For example, the static points identified by radar scene flow can be used to estimate the vehicle odometry. The consistency between this estimated odometry and the observed odometry from the co-located GPS/INS on the vehicle forms a natural constraint and can be used to supervise scene flow estimation. Similar consistency constraints can be also found for the optical flow estimated from co-located cameras and the projected radar scene flow onto the image plane.

While the aforementioned examples are intuitive, retrieving accurate supervision signals from co-located sensors is non-trivial. With the same example of optical and scene flow consistency, minimizing flow errors on the image plane suffers from the depth-unaware perspective projection, potentially incurring weaker constraints to the scene flow of far points. This motivates the following research question: *How to retrieve the cross-modal supervision signals from co-located sensors on a vehicle and apply them collectively to bootstrap radar scene flow learning.*

Towards answering it, in this work we consider *opportunistically* exploiting useful supervision signals from three commonly co-existent sensors with the 4D radar on a vehicle: odometer (GPS/INS), LiDAR, and RGB camera (See Fig. 1). This cross-modal supervision is expected to help us realize radar scene flow learning without human annotation. In our setting, the multi-modal data are only available in the training phase, and only 4D radar is used during the inference stage. Our contributions can be summarized as follows:

- Our work is the first 4D radar scene flow learning using cross-modal supervision from co-located heterogeneous sensors on an autonomous vehicle.
- We introduce a multi-task model architecture for the identified cross-modal learning problem and propose loss functions to effectively engage scene flow estimation using multiple cross-modal constraints for model training.
- We demonstrate the state-of-the-art performance of the proposed CMFlow method on a public dataset and show its effectiveness in downstream tasks as well.

2. Related Work

Scene flow. Scene flow was first defined in [44] as a 3D uplift of optical flow that describes the 3D displacement of points in the scene. Traditional approaches resolve pixel-wise scene flow from either RGB or RGB-D images based on prior knowledge assumptions [6, 13–15, 38, 39, 45, 48] or by training deep networks in a supervised [17, 18, 28, 29, 37] or unsupervised [16, 26, 53, 55] way. In contrast, some other methods directly infer point-wise scene flow from 3D sparse point clouds. Among them, some methods rely on online optimization to solve scene

flow [8, 25, 33]. Recently, inspired by the success of point cloud feature learning [35, 36, 40, 51], deep learning-based methods [3, 12, 27, 52] have been dominant for point cloud-based scene flow estimation.

Deep scene flow on point clouds. Current point cloud-based scene flow estimation methods [2, 3, 10–12, 21, 24, 27, 31, 34, 46, 47, 49, 52] established state-of-the-art performance by leveraging large amount of data for training. Many of them [3, 34, 46, 47, 49] learn scene flow estimation in a fully-supervised manner with ground truth flow. These methods, albeit showing promising results, demand scene flow annotations, of which the acquisition is labour-intensive and costly. Another option is to leverage a simulated dataset [29] for training, yet this may result in poor generalization when applied to the real data. To avoid both human labour and the pitfalls of synthetic data, some methods design self-supervised learning frameworks [2, 21, 24, 31, 52] that exploit supervision signals from the input data. Compared with their supervised counterparts, these methods require no annotated labels and thus can train their models on unannotated datasets. However, the performance of these self-supervised learning methods is limited [21, 52] by the fact that no real labels are used to supervise their models. Some recent methods [10, 11] try to seek a trade-off between annotation efforts and performance by combining the ego-motion and manually annotated background segmentation labels. Although *pseudo* ground truth ego-motion can be easily accessed from onboard odometry sensors (GPS/INS), annotating background mask labels is still expensive and need human experts to identify foreground entities from complex scenarios.

Radar scene flow. Previous works mostly estimate scene flow on dense point clouds captured by LiDAR or rendered from stereo images. Thus, they cannot be directly extended to the sparse and noisy radar point clouds. To fill the gap, a recent work [9] proposes a self-supervised pipeline for radar scene flow estimation. However, just like in other self-supervised methods, the lack of real supervision signals limits its scene flow estimation performance and thus hinders its application to more downstream tasks.

Unlike the aforementioned works, we propose to retrieve supervision signals from co-located sensors in an automatic manner without resorting to any human intervention during training. Note that we use cross-modal data to provide all supervision at once only in the training stage and do not require other modalities during inference.

3. Method

3.1. Problem Definition

Scene flow estimation aims to solve a motion field that describes the non-rigid transformations induced both by the motion of the ego-vehicle and the dynamic objects in the

scene. For point cloud-based scene flow, the inputs are two consecutive point clouds, the source one $\mathbf{P}^s = \{\mathbf{p}_i^s = \{\mathbf{c}_i^s, \mathbf{x}_i^s\}\}_{i=1}^N$ and the target one $\mathbf{P}^t = \{\mathbf{p}_i^t = \{\mathbf{c}_i^t, \mathbf{x}_i^t\}\}_{i=1}^M$, where $\mathbf{c}_i^s, \mathbf{c}_i^t \in \mathbb{R}^3$ are the 3D coordinates of each point, and $\mathbf{x}_i^s, \mathbf{x}_i^t \in \mathbb{R}^C$ are their associated raw point features. The outputs are point-wise 3D motion vectors $\mathbf{F} = \{\mathbf{f}_i \in \mathbb{R}^3\}_{i=1}^N$ that align each point in $\mathbf{P}^s$ to its corresponding position $\mathbf{c}_i' = \mathbf{c}_i^s + \mathbf{f}_i$ in the target frame. Note that $\mathbf{P}^s$ and $\mathbf{P}^t$ do not necessarily have the same number of points and there is no strict point-to-point correspondence between them under real conditions. Therefore, the corresponding location $\mathbf{c}_i'$ is not required to coincide with any points in the target point cloud $\mathbf{P}^t$. In our case of 4D radar, the raw point features $\mathbf{x}_i^s, \mathbf{x}_i^t \in \mathbb{R}^2$ include the relative radial velocity (RRV) and the radar cross-section (RCS) measurements [1]. RRV measurements, resolved by analyzing Doppler shift in radar frequency data, contain partial motion information of points. RCS can be seen as the proxy reflectivity of each point, which is mainly affected by the reflection property of the target and the incident angle of the beams.

3.2. Overview

The overall pipeline of our proposed method is depicted in Fig. 2. To bootstrap cross-modal supervised learning, we apply a multi-task model that predicts radar scene flow in two stages. The first stage starts by extracting base features with two input point clouds, which are then forwarded to two independent heads to infer per-point moving probabilities and initial point-wise scene flow vectors. On top of them, the second stage first infers a rigid transformation respective to radar ego-motion and outputs a binary motion segmentation mask. Then, the final scene flow is obtained by refining the flow vectors of identified static points with the derived rigid transformation. In summary, our multi-task model's outputs include a rigid transformation (*i.e.* the ego-motion), a motion segmentation mask (*i.e.* which targets are static or dynamic), and the final refined scene flow.

To supervise these predictions, we extract corresponding supervision signals from co-located modalities and train the entire model end-to-end by minimizing a loss $\mathcal{L}$ composed of three terms:

$$\mathcal{L} = \mathcal{L}_{ego}^{\bullet} + \mathcal{L}_{seg}^{\bullet} + \mathcal{L}_{flow}^{\bullet}. \quad (1)$$

Here, $\mathcal{L}_{ego}^{\bullet}$ is the ego-motion error that supervises the rigid transformation estimation using the odometry information. Our motion segmentation error $\mathcal{L}_{seg}^{\bullet}$ constrains the predicted moving probabilities with a point-wise pseudo motion segmentation label, which is obtained by fusing information from the odometer and LiDAR. We further supervise the final scene flow with signals given by LiDAR and RGB camera in $\mathcal{L}_{flow}^{\bullet}$. In the following section (Sec. 3.3), we first briefly introduce each module of our model. We then explain how we extract signals from co-located modal-

ities to supervise our outputs in the training phase (Sec. 3.4). More details on our method can be found in the supplementary materials.

3.3. Model Architecture

Similar to [2, 9–11, 43], our model is designed in a two-stage fashion, with a rough initial flow derived in the first stage and refined in the second stage to obtain the final estimate of scene flow, as shown in Fig. 2. A description of all components of our model can be found below.

Backbone. Following [9, 10, 27, 52], our backbone network directly operates on unordered point clouds $\mathbf{P}^s$ and $\mathbf{P}^t$ to encode point-wise latent features. In particular, we first apply the *set conv* layers [27] to robustly extract multi-scale local features for individual point clouds and propagate features from $\mathbf{P}^t$ to $\mathbf{P}^s$ using the *cost volume* layer [52] for feature correlation. We then concatenate multi-stage features of $\mathbf{P}^s$ (including the correlated features) and forward them into another multi-scale *set conv* layers to generate the base backbone features $\mathbf{E} \in \mathbb{R}^{N \times C_e}$. Specifically, the max-pooling operation is used along the channel axis to restore the global scene features after each *set conv* layer, which are then concatenated to per-point local features.

Initial flow and motion segmentation head. Given the base backbone features $\mathbf{E}$ that represents the inter-frame motion and intra-frame local-global information for each point $\mathbf{p}_i^s \in \mathbf{P}^s$, we apply two task-specific heads for decoding. The first head is to produce an initial scene flow $\hat{\mathbf{F}}^{init} = \{\hat{\mathbf{f}}_i^{init} \in \mathbb{R}^3\}_{i=1}^N$, while another is for generating a probability map $\hat{\mathbf{S}} = \{\hat{s}_i \in [0, 1]\}_{i=1}^N$ that denotes the moving probabilities of points in $\mathbf{P}^s$ (*w.r.t.* the world frame). We implement both heads with a multi-layer perceptron and map the output to probabilities with *Sigmoid*($\cdot$) in the motion segmentation head.

Ego-motion head. With the natural correspondences $\{\mathbf{c}_i^s, \mathbf{c}_i^s + \hat{\mathbf{f}}_i^{init}\}_{i=1}^N$ formed by the initial scene flow $\hat{\mathbf{F}}^{init}$ between two frames and the probability map $\hat{\mathbf{S}}^1$, we restore a rigid transformation $\hat{\mathbf{T}} \in \mathbb{R}^{4 \times 4}$ that describes the radar ego-motion using the differentiable weighted Kabsch algorithm [19]. To mitigate the impact of the flow vectors from moving points, we compute $1 - \hat{\mathbf{S}}$ as our weights and normalize the sum of them to 1. Besides, we generate a binary motion segmentation mask by thresholding $\hat{\mathbf{S}}$ with a fixed value η_b to indicate the moving status of each point. We use this binary mask as our motion segmentation output and to identify stationary points for flow refinement below.

Refinement layer. As the flow vectors of static points are only caused by the radar's ego-motion, we can regularize their initial predictions with the more reliable rigid transformation $\hat{\mathbf{T}}$. The refinement operation is simply replacing the initial flow vector $\hat{\mathbf{f}}_i^{init}$ of identified stationary points

¹Specifically, we use the pseudo label $\tilde{\mathbf{S}}$ (c.f. Sec. 3.4) instead of $\hat{\mathbf{S}}$ in the ego-motion head during training for stable scene flow learning.

time duration between two frames. Then, we compensate the radar ego-motion and get the object absolute radial velocity $\Delta v_i = \text{abs}(v_i - v_i^r)$. With per-point Δv_i , the pseudo motion segmentation label $\mathbf{S}^v = \{s_i^v \in \{0, 1\}\}_{i=1}^N$ can be derived by thresholding, where 1 denotes moving points. More details on our thresholding strategy can be found in the supplementary materials. Note that one shortcoming is that tangentially moving targets are not distinguished.

Besides the odometry information, we also leverage the LiDAR data to generate a pseudo foreground (movable objects) segmentation label $\mathbf{S}^{fg}$. To this end, we first feed LiDAR point clouds into an off-the-shelf 3D multi-object tracking (MOT) pretrained model [50]. Then we divide radar points from $\mathbf{P}^s$ into the foreground and background using the bounding boxes (e.g. pedestrian, car, cyclist) produced by 3D MOT to create $\mathbf{S}^{fg}$. Besides, we can also assign pseudo scene flow vector label to foreground points by: a) retrieving the ID for each bounding box from the first frame, b) computing the inter-frame transformation for them, c) deriving the translation vector for each in-box point based on the assumption that all points belonging to the same object share a *universal* rigid transformation. The resulting pseudo scene flow label is denoted as $\mathbf{F}^{fg} = \{\mathbf{f}_i^{fg}\}_{i=1}^N$, where we leave the label empty for all identified background points.

Given the two segmentation labels $\mathbf{S}^v$ and $\mathbf{S}^{fg}$, directly fusing them is impeded by their domain discrepancy, i.e., not all foreground points are moving. Therefore, we propose to distill the moving points from $\mathbf{S}^{fg}$ by discarding foreground points that either keep still or move very slowly. We implement this by removing the rigid flow component $\mathbf{F}^r$ from $\mathbf{F}^{fg}$ and get the non-rigid flow component of all foreground points. Then we obtain a new pseudo motion segmentation label $\mathbf{S}^l = \{s_i^l \in \{0, 1\}\}_{i=1}^N$ by classifying points with apparent non-rigid flow as dynamic. A more reliable pseudo motion segmentation $\mathbf{S} = \{s_i\}_{i=1}^N$ can be consequently obtained by fusing $\mathbf{S}^l$ and $\mathbf{S}^v$. For points classified as moving in $\mathbf{S}^l$, we have high confidence about their status and thus label these points as moving in $\mathbf{S}$. For the rest of the points, we label their s_i according to $\mathbf{S}^v$. Finally, as seen in Fig. 2, our motion segmentation loss can be formulated by encouraging the estimated moving probabilities $\hat{\mathbf{S}}$ to be close to the pseudo motion segmentation label $\mathbf{S}$:

$$\mathcal{L}_{seg}^{\bullet} = \frac{1}{2} \left(\frac{\sum_{i=1}^N (1 - s_i) \log(1 - \hat{s}_i)}{\sum_{i=1}^N (1 - s_i)} + \frac{\sum_{i=1}^N s_i \log(\hat{s}_i)}{\sum_{i=1}^N s_i} \right). \quad (3)$$

Here, we use the average loss of static and moving losses to address the class imbalance issue.

Scene flow loss. In the final scene flow output $\hat{\mathbf{F}}$, the flow vectors of static points have been implicitly supervised by the ego-motion loss (c.f. Eq. (2)). In order to further constrain the flow vectors of moving points, we formulate two new loss functions. The first one is $\mathcal{L}_{mot}^{\bullet}$, which is based

on the pseudo scene flow label $\mathbf{F}^{fg}$ derived from 3D MOT results. In this loss, we only constrain the flow vectors of those moving points identified in $\mathbf{S}^l$ through:

$$\mathcal{L}_{mot}^{\bullet} = \frac{1}{\sum_{i=1}^N s_i^l} \sum_{i=1}^N \left\| s_i^l (\hat{\mathbf{f}}_i - \mathbf{f}_i^{fg}) \right\|_2. \quad (4)$$

In addition to utilizing the odometer and LiDAR for cross-modal supervision, we also propose to extract supervision signals from the RGB camera. Specifically, we formulate a loss $\mathcal{L}_{opt}^{\bullet}$ using pseudo optical flow labels $\mathbf{W} = \{\mathbf{w}_i \in \mathbb{R}^2\}_{i=1}^N$. To get this pseudo label, we first feed synchronized RGB images into a pretrained optical flow estimation model [42] to produce an optical flow image. Then, we project the coordinate $\mathbf{c}_i^s$ of each point onto the image plane and draw the optical flow vector at the corresponding pixel $\mathbf{m}_i$ of each point. Given $\mathbf{W}$, it is intuitive to construct the supervision for our scene flow prediction by projecting it on the image plane. However, minimizing the flow divergence in pixel scale has less impact for far radar points due to the depth-unawareness during perspective projection. Instead, we directly take the point-to-ray distance as the training objective, which is more insensitive to points at different ranges. The loss function can be written as:

$$\mathcal{L}_{opt}^{\bullet} = \frac{1}{\sum_{i=1}^N s_i} \sum_{i=1}^N s_i \mathcal{D}(\mathbf{c}_i^s + \hat{\mathbf{f}}_i, \mathbf{m}_i + \mathbf{w}_i, \theta) \quad (5)$$

where $\mathcal{D}$ denotes the operation that computes the distance between the warped point $\mathbf{c}_i^s + \hat{\mathbf{f}}_i$ and the ray traced from the warped pixel $\mathbf{m}_i + \mathbf{w}_i$. θ denotes sensor calibration parameters. Note that we only consider the scene flow of moving points identified in $\mathbf{S}$ here. Apart from the above two loss functions used to constrain the final scene flow, we also employ the self-supervised loss $\mathcal{L}_{self}^{\bullet}$ in [9] to complement our cross-modal supervision. See the supplementary for more details. The overall scene flow loss is formulated as:

$$\mathcal{L}_{flow}^{\bullet} = \mathcal{L}_{mot}^{\bullet} + \lambda_{opt} \mathcal{L}_{opt}^{\bullet} + \mathcal{L}_{self}^{\bullet}, \quad (6)$$

where we set the weight $\lambda_{opt} = 0.1$ in all our experiments.

4. Experiments

4.1. Experimental Setup

Dataset. For our experiments, we use the *View-of-Delft* (VoD) dataset [32], which provides synchronized and calibrated data captured by co-located sensors, including a 64-beam LiDAR, an RGB camera, RTK-GPS/IMU based odometer and a 4D radar sensor. As is often the case with datasets focused on object recognition, the test set annotations of the official VoD dataset are withheld for benchmarking. However, our task requires custom scene flow metrics for which we need annotations to generate ground truth labels. Therefore, we divide new splits ourselves from

Method	Sup.	EPE [m]↓	AccS↑	AccR↑	RNE [m]↓	MRNE [m]↓	SRNE [m]↓
ICP [4]	None	0.344	0.019	0.106	0.138	0.148	0.137
Graph Prior* [33]	None	0.445	0.070	0.104	0.179	0.186	0.176
JGWTf* [31]	Self	0.375	0.022	0.103	0.150	0.139	0.151
PointPWC [52]	Self	0.422	0.026	0.113	0.169	0.154	0.170
FlowStep3D [21]	Self	0.292	0.034	0.161	0.117	0.130	0.115
SLIM* [2]	Self	0.323	0.050	0.170	0.130	0.151	0.126
RaFlow [9]	Self	0.226	0.190	0.390	0.090	0.114	0.087
CMFlow	Cross	0.141	0.233	0.499	0.057	0.073	0.054
CMFlow (T)	Cross	0.130	0.228	0.539	0.052	0.072	0.049

Table 1. Scene flow evaluation results on the *Test* set. The mean metric values across all test frames are reported. * indicates that we reproduce these methods referring to original papers since their source codes are not public. The best results on each metric are shown in **bold**. ↑ means bigger values are better, and vice versa.

the official sets (*i.e.*, *Test*, *Val*, *Train*) to support our evaluation. Given sequences of data frames, we form pairs of consecutive radar point clouds as our scene flow samples for training and inference. We only generate ground truth scene flow and motion segmentation labels for samples from our *Val*, *Test* sets and leave the *Train* set unlabelled as our method has no need for ground truth scene flow annotations for training. Please see the supplementary details on our dataset separation and labelling process.

Metrics. We use three standard metrics [27, 31, 52] to evaluate different methods on scene flow estimation, including a) *EPE* [m]: average end-point-error (L_2 distance) between ground truth scene flow vectors and predictions, b) *AccS/AccR*: the ratio of points that meet a strict/relaxed condition, *i.e.* $EPE < 0.05/0.1$ m or the relative error $< 5\%/10\%$. We also use the c) *RNE* [m] metric [9] that computes resolution-normalized *EPE* by dividing *EPE* by the ratio of 4D radar and LiDAR resolution. This can induce sensor-specific consideration and maintain a fair comparison between different sensors. Besides, we compute the *RNE* [m] for moving points and static points respectively, and denote them as *MRNE* [m] and *SRNE* [m].

Baselines. For overall comparison, we apply seven state-of-the-art methods as our baselines, including five self-supervised learning-based methods [2, 9, 21, 31, 52] and two non-learning-based methods [4, 33], as seen in Tab. 1. To keep a fair comparison, we use their default hyperparameter settings. We retrain scene flow models offline on our VoD *Train* set for learning-based methods and directly optimize scene flow results online for non-learning-based ones.

Implementation details. We use the Adam optimizer [20] to train all models in our experiments. The learning rate is initially set as 0.001 and exponentially decays by 0.9 per epoch. The *Val* set is used for both model selection during training and for determining the values of our hyperparameters. We set our classification threshold $\eta_b = 0.5$ (c.f. Sec. 3.3) for all experiments. When activating our temporal update module, the length of mini-clips T is set as 5. Please refer to the supplementary materials for our hyperparameter searching process.

	O	L	C	EPE [m]↓	AccS↑	AccR↑	RNE [m]↓
(a)				0.228	0.184	0.392	0.091
(b)	✓			0.161	0.203	0.442	0.065
(c)	✓	✓		0.145	0.228	0.482	0.058
(d)	✓		✓	0.159	0.216	0.458	0.064
(e)	✓	✓	✓	0.141	0.233	0.499	0.057

Table 2. Ablation experiments on combing supervision signals from diverse modalities. Abbreviations: odometer (O), LiDAR (L), camera (C). Note that we disable the temporal update scheme in this study to highlight the impact of modalities for training.

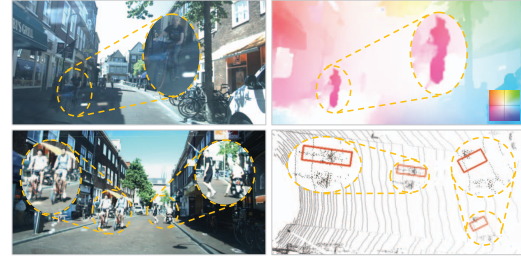


Figure 3. Illustration of the causes of noisy supervision signals from camera and LiDAR. The top row shows an example of the noisy optical flow estimation on RGB images. The bottom row exhibits unreliable object recognition on LiDAR point clouds. We enlarge regions of interest and mark them with orange circles.

4.2. Scene Flow Evaluation

Overall results. We quantitatively compare the performance of our methods to baselines on the *Test* set, as shown in Tab. 1. Compared with both non-learning-based and self-supervised counterparts, CMFlow shows remarkable improvement on all metrics by leveraging cross-modal supervision signals for training. Our method outperforms the second-best approach [9] by 37.6% on EPE, implying drastically more reliability. Our performance is further improved when adding the temporal update scheme (c.f. Sec. 3.3) in the backbone, *i.e.* CMFlow (T). We also observe that the performance slightly degrades on the AccS metric when activating the temporal update. This suggests us that introducing temporal information will somewhat disturb the original features and thus reduce the fraction of already fine-grained (*e.g.* $EPE < 0.05$ m) predictions.

Impact of modalities. A key mission of this work is to investigate how supervision signals from different modalities help our radar scene flow estimation. To analyze their impact, we conduct ablation studies and show the breakdown results in Tab. 2. The results show that the odometer leads to the biggest performance gain for our method (c.f. row (b)). This can be credited to the fact, that most measured points are static (*e.g.* 90.5% for the *Test* set) and their flow predictions can be supervised well by the ego-motion supervision (Eq. (2)). Moreover, the odometer guides the generation of pseudo motion segmentation labels (Sec. 3.4), which is indispensable for our two-stage model architecture.

Based on row (b), both LiDAR and camera contribute

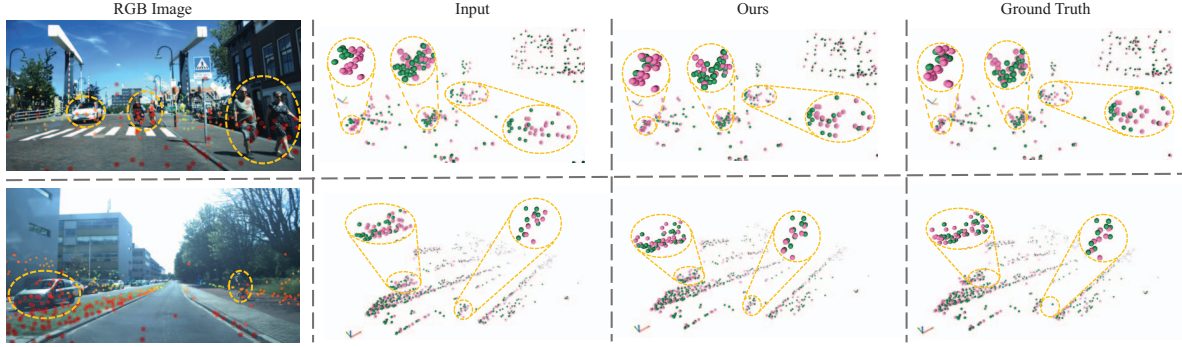


Figure 4. Qualitative scene flow results in two scenes. From left to right: 1) radar points from the source frame projected to the corresponding RGB image (points are coloured by distance from the sensor), 2) two input radar point clouds, the source one (pink) and the target one (green), 3) the source point cloud warped by our predicted scene flow and the target radar point cloud, 4) the source point cloud warped by ground truth scene flow and the target one. We mark dynamic objects in amber and apply the zooming-in operation for them.

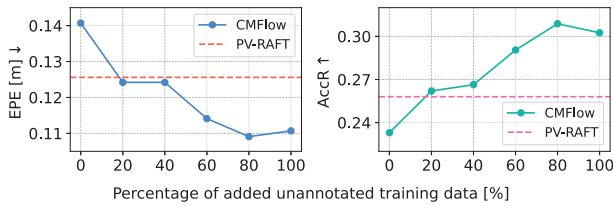


Figure 5. Analysis of the performance when adding more unannotated training data. For the training of CMFlow, we retain the samples from the *Train* set and add different percentages of data from the extra unannotated VoD part, which provides $\sim 28.5k$ more training samples.

to a further improvement on all metrics, as seen in row (c)-(e). These results validate that our method can effectively exploit useful supervision signals from each modality to facilitate better scene flow estimation. However, compared to that from the odometer, the gains brought by these two sensors are smaller, especially for the camera, which only increases the AccS by 1.3%. In our opinion, the reason for this is two-fold. First, in Eq. (4) and Eq. (5), the pseudo scene/optical flow labels only constrain the flow vectors of identified moving points, which are significantly fewer than static ones. Thus, they have limited influence on the overall performance. Second, the supervision signals extracted from these sensor modalities could be noisy, as exhibited in Fig. 3. The dashboard reflections on the car windscreen severely disturb the optical flow estimation on images, which further results in erroneous supervision in Eq. (5). As for LiDAR point clouds, both inaccurate object localization and false negative detections occur in the 3D MOT results, which not only affect the generation of pseudo motion segmentation labels (Eq. (3)) but bring incorrect constraints to scene flow in Eq. (4).

Impact of the amount of unannotated data. Being able to use more unannotated data for training is an inherent advantage of our cross-modal supervised method compared to fully-supervised ones as no additional annotation efforts

	Label S^v	Label S^l	A.D.	mIoU (%)	Gain (%)
(a)				46.9	-
(b)	✓			52.8	+5.9
(c)	✓	✓		54.1	+1.3
(d)	✓	✓	✓	57.1	+3.0

Table 3. Motion segmentation evaluation. For row (a), we report the results of the self-supervised baseline [9]. In row (b), we replace S by S^v provided by odometer. A.D. denotes adding extra unannotated data (c.f. Fig. 5) for training.

are required. Here, we are interested in if the performance of CMFlow could surpass that of fully-supervised methods when more unannotated data is available for training. To this end, we further mix in an extra amount of unlabeled data provided by the VoD dataset [32]² in our cross-modal training sets. As for fully-supervised methods, we select the state-of-the-art method, PV-RAFT [49], for comparison. As PV-RAFT needs ground truth scene flow annotations for training, we utilize all available annotated samples from the *Train* set for it. The analysis of the correlation between the percentage of added unannotated training data and the performance is shown in Fig. 5. As we can see, the performance of CMFlow improves by a large margin on both two metrics by using extra training data. In particular, after adding only 20% of the extra unannotated training samples ($\sim 140\%$ more than the number of training samples used for PV-RAFT), CMFlow can already outperform PV-RAFT [49] trained with less annotated samples. This implies the promise of our method for utilizing a large amount of unannotated data in the wild.

Qualitative results. In Fig. 4, we showcase example scene flow results of CMFlow (trained with the extra training samples) compared to the ground truth scene flow. By applying the estimated scene flow (*i.e.* moving each point in the

²This part is currently in a beta testing phase, only available for selected research groups. Except for having no object annotations, it provides the same modalities of input data as the official one and has $\sim 28.5k$ frames.

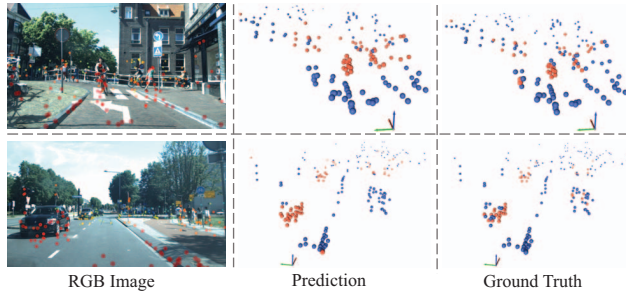


Figure 6. Visualization of motion segmentation. The left column shows the corresponding image with radar points (coloured by range) projected onto it. In the middle and right columns, moving points are shown in orange while static points are shown in blue.

source frame by its estimated 3D motion vector), both the static background and multiple objects with different motion patterns are aligned well between two frames. It can also be observed that our method can give accurate scene flow predictions, close to the ground truth one.

4.3. Subtask Evaluation

Motion segmentation evaluation. Apart from scene flow estimation, our multi-task model can additionally predict a motion segmentation mask that represents the real moving status of each radar point (c.f. Sec. 3.3). Here, we evaluate this prediction of CMFlow and analyze the impact of its performance in Tab. 3. Since this is a binary classification task for each point, the mean intersection over union (mIoU) is computed by taking the IoU of moving and static classes and averaging them. As the two ingredients to form the final pseudo motion segmentation label $\mathbf{S}$ used to supervise our output in Eq. (3), both $\mathbf{S}^v$ and $\mathbf{S}^l$ contributes to our performance improvement on the motion segmentation task (row (a)-(c)). Moreover, our mIoU further increases with extra training data to learn motion segmentation on 4D radar point clouds. We also visualize some qualitative motion segmentation results of row (d) in Fig. 6, where our method can segment moving points belonging to multiple dynamic objects accurately in complicated scenarios.

Ego-motion estimation evaluation. One important feature of our method is that we can estimate a rigid transformation that represents relative radar ego-motion transform between two consecutive frames in dynamic scenes (c.f. Sec. 3.3). To demonstrate this feature, we evaluate this output on the *Test* set and show the ablation study results in Tab. 4 with two metrics: the relative translation error (RTE) and the relative angular error (RAE). With the cross-modal supervision from the odometry in row (b), we can directly constrain our ego-motion estimation in Eq. (2) and thus improve the performance by a large margin. Using LiDAR and camera supervision (c.f. row (c)) can also help as they lead to better motion segmentation and scene flow outputs, which further benefit the compactly associated ego-motion

	O	L + C	A.D.	T	RTE [m]	RAE [°]
(a)					0.090	0.336
(b)	✓				0.086	0.183
(c)	✓	✓			0.085	0.145
(d)	✓	✓	✓		0.071	0.089
(e)	✓	✓	✓	✓	0.066	0.090

Table 4. Evaluation of ego-motion estimation between two frames. T denotes we activate the temporal update module.

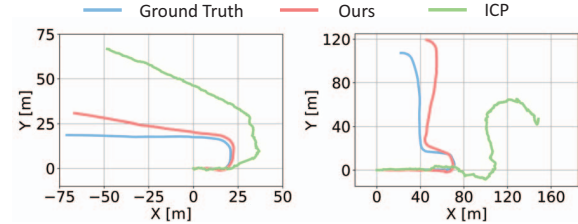


Figure 7. Odometry results as a byproduct of our scene flow estimation. The ground truth is generated using the RTK-GPS/IMU measurements. We plot the results on two challenging test sequences. Please refer to more results in supplementary videos.

estimation. We also activate the temporal update module in the backbone, which also increases the overall performance. With our high-level accuracy on ego-motion estimation between consecutive frames, we are also interested in whether our results can be used for the more challenging long-term odometry task. We accumulate the inter-frame transformation estimations and plot two ego-vehicle trajectories in Fig. 7. Without any global optimization, our method can provide accurate long-term trajectory estimation in dynamic scenes by only estimating inter-frame transformations and remarkably outperform the ICP baseline [4].

5. Conclusion

In this paper, we presented a novel cross-modal supervised approach, CMFlow, for estimating 4D radar scene flows. CMFlow is unique in that it does not require manual annotation for training. Instead, it uses complementary supervision signals retrieved from co-located heterogeneous sensors, such as odometer, LiDAR and camera, to constrain the outputs from our multi-task model. Our experiments show that CMFlow outperforms our baseline methods in all metrics and can surpass the fully-supervised method when sufficient unannotated samples are used in our training. CMFlow can also improve two downstream tasks, *i.e.*, motion segmentation and ego-motion estimation. We hope our work will inspire further investigation of cross-modal supervision for scene flow estimation and its application to more downstream tasks, such as multi-object tracking and point cloud accumulation.

Acknowledgment. This research is supported by the EPSRC, as part of the CDT in Robotics and Autonomous Systems hosted at the Edinburgh Centre of Robotics (EP/S023208/1), and the Google Cloud Research Credits program with the award GCP19980904.

References

- [1] David K Barton. *Radar system analysis and modeling*. Artech House, 2004. 3
- [2] Stefan Andreas Baur, David Josef Emmerichs, Frank Moosmann, Peter Pinggera, Björn Ommer, and Andreas Geiger. SLIM: Self-Supervised LiDAR Scene Flow and Motion Segmentation. In *CVPR*, pages 13126–13136, 2021. 1, 2, 3, 6
- [3] Aseem Behl, Despoina Paschalidou, Simon Donné, and Andreas Geiger. PointFlowNet: Learning Representations for Rigid Motion Estimation From Point Clouds. In *CVPR*, pages 7954–7963, 2019. 2
- [4] P.J. Besl and Neil D. McKay. A method for registration of 3-D shapes. *PAMI*, 14(2):239–256, 1992. 6, 8
- [5] Stefan Bricken, Florian Ruf, and Felix Höhne. Recent evolution of automotive imaging radar and its information content. *IET Radar Sonar Navig.*, 12(10):1078–1081, 2018. 1
- [6] Jan Čech, Jordi Sanchez-Riera, and Radu Horaud. Scene flow estimation by growing correspondence seeds. In *CVPR*, pages 3129–3136, 2011. 2
- [7] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *SSST*, pages 103–111, 2014. 4
- [8] Ayush Dewan, Tim Caselitz, Gian Diego Tipaldi, and Wolfram Burgard. Rigid scene flow for 3D LiDAR scans. In *IROS*, pages 1765–1770, 2016. 2
- [9] Fangqiang Ding, Zhijun Pan, Yimin Deng, Jianning Deng, and Chris Xiaoxuan Lu. Self-Supervised Scene Flow Estimation With 4-D Automotive Radar. *RA-L*, pages 1–8, 2022. 1, 2, 3, 5, 6, 7
- [10] Guanting Dong, Yueyi Zhang, Hanlin Li, Xiaoyan Sun, and Zhiwei Xiong. Exploiting Rigidity Constraints for LiDAR Scene Flow Estimation. In *CVPR*, pages 12776–12785, 2022. 1, 2, 3
- [11] Zan Gojcic, Or Litany, Andreas Wieser, Leonidas J Guibas, and Tolga Birdal. Weakly Supervised Learning of Rigid 3D Scene Flow. In *CVPR*, pages 5692–5703, 2021. 1, 2, 3
- [12] Xiuye Gu, Yijie Wang, Chongruo Wu, Yong Jae Lee, and Panqu Wang. HPLFlowNet: Hierarchical Permutohedral Lattice FlowNet for Scene Flow Estimation on Large-Scale Point Clouds. In *CVPR*, pages 3254–3263, 2019. 1, 2
- [13] Simon Hadfield and Richard Bowden. Kinecting the dots: Particle based scene flow from depth sensors. In *ICCV*, pages 2290–2295, 2011. 2
- [14] Michael Hornacek, Andrew Fitzgibbon, and Carsten Rother. SphereFlow: 6 DoF scene flow from RGB-D pairs. In *CVPR*, pages 3526–3533, 2014. 2
- [15] Frédéric Hugué and Frédéric Devernay. A variational method for scene flow estimation from stereo sequences. In *CVPR*, pages 1–7, 2007. 2
- [16] Junhwa Hur and Stefan Roth. Self-supervised monocular scene flow estimation. In *CVPR*, pages 7396–7405, 2020. 2
- [17] Eddy Ilg, Tonmoy Saikia, Margret Keuper, and Thomas Brox. Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation. In *ECCV*, pages 614–630, 2018. 2
- [18] Huaizu Jiang, Deqing Sun, Varun Jampani, Zhaoyang Lv, Erik Learned-Miller, and Jan Kautz. Sense: A shared encoder network for scene-flow estimation. In *ICCV*, pages 3195–3204, 2019. 2
- [19] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A*, 32(5):922–923, 1976. 3
- [20] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. 6
- [21] Yair Kittenplon, Yonina C Eldar, and Dan Raviv. Flow-Step3D: Model Unrolling for Self-Supervised Scene Flow Estimation. In *CVPR*, pages 4114–4123, 2021. 1, 2, 6
- [22] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, pages 12697–12705, 2019. 1
- [23] Ruibo Li, Guosheng Lin, Tong He, Fayao Liu, and Chunhua Shen. HCRF-Flow: Scene Flow from Point Clouds with Continuous High-order CRFs and Position-aware Flow Embedding. In *CVPR*, pages 364–373, 2021. 1
- [24] Ruibo Li, Chi Zhang, Guosheng Lin, Zhe Wang, and Chunhua Shen. RigidFlow: Self-Supervised Scene Flow Learning on Point Clouds by Local Rigidity Prior. In *CVPR*, pages 16959–16968, 2022. 1, 2
- [25] Xueqian Li, Jhony Kaesemodel Pontes, and Simon Lucey. Neural scene flow prior. *NIPS*, 34:7838–7851, 2021. 2
- [26] Liang Liu, Guangyao Zhai, Wenlong Ye, and Yong Liu. Unsupervised Learning of Scene Flow Estimation Fusing with Local Rigidity. In *IJCAI*, pages 876–882, 2019. 2
- [27] Xingyu Liu, Charles R Qi, and Leonidas J Guibas. FlowNet3D: Learning Scene Flow in 3D Point Clouds. In *CVPR*, pages 529–537, 2019. 1, 2, 3, 6
- [28] Zhaoyang Lv, Kihwan Kim, Alejandro Troccoli, Deqing Sun, James M Rehg, and Jan Kautz. Learning Rigidity in Dynamic Scenes with a Moving Camera for 3D Motion Field Estimation. In *ECCV*, pages 468–484, 2018. 2
- [29] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, pages 4040–4048, 2016. 2
- [30] Michael Meyer and Georg Kuschik. Automotive Radar Dataset for Deep Learning Based 3D Object Detection. In *EuRAD*, pages 129–132, 2019. 1
- [31] Himangi Mittal, Brian Okorn, and David Held. Just go with the flow: Self-supervised scene flow estimation. In *CVPR*, pages 11177–11185, 2020. 1, 2, 6
- [32] Andras Palffy, Ewoud Pool, Srimannarayana Baratam, Julian FP Kooij, and Dariu M Gavrilă. Multi-class Road User Detection with 3+ 1D Radar in the View-of-Delft Dataset. *RA-L*, 7(2):4961–4968, 2022. 1, 5, 7
- [33] Jhony Kaesemodel Pontes, James Hays, and Simon Lucey. Scene flow from point clouds with or without learning. In *3DV*, pages 261–270, 2020. 2, 6
- [34] Gilles Puy, Alexandre Boulch, and Renaud Marlet. Flot: Scene flow on point clouds guided by optimal transport. In *ECCV*, pages 527–544, 2020. 1, 2

- [35] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *CVPR*, pages 652–660, 2017. 1, 2
- [36] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NIPS*, 30, 2017. 1, 2
- [37] Yi-Ling Qiao, Lin Gao, Yu-Kun Lai, Fang-Lue Zhang, Mingzhe Yuan, and Shihong Xia. SF-Net: Learning Scene Flow from RGB-D Images with CNNs. In *BMVC*, page 281, 2018. 2
- [38] Julian Quiroga, Thomas Brox, Frédéric Devernay, and James Crowley. Dense Semi-rigid Scene Flow Estimation from RGBD Images. In *ECCV*, pages 567–582, 2014. 2
- [39] Julian Quiroga, Frédéric Devernay, and James Crowley. Local/global scene flow estimation. In *ICIP*, pages 3850–3854, 2013. 2
- [40] Hang Su, Varun Jampani, Deqing Sun, Subhransu Maji, Evangelos Kalogerakis, Ming-Hsuan Yang, and Jan Kautz. Splatnet: Sparse lattice networks for point cloud processing. In *CVPR*, pages 2530–2539, 2018. 1, 2
- [41] Shunqiao Sun and Yimin D Zhang. 4D automotive radar sensing for autonomous vehicles: A sparsity-oriented approach. *JSTSP*, 15(4):879–891, 2021. 1
- [42] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419. Springer, 2020. 5
- [43] Ivan Tishchenko, Sandro Lombardi, Martin R Oswald, and Marc Pollefeys. Self-supervised learning of non-rigid residual flow and ego-motion. In *3DV*, pages 150–159, 2020. 3
- [44] Sundar Vedula, Peter Rander, Robert Collins, and Takeo Kanade. Three-dimensional scene flow. *PAMI*, 27(3):475–480, 2005. 2
- [45] Christoph Vogel, Konrad Schindler, and Stefan Roth. Piecewise rigid scene flow. In *ICCV*, pages 1377–1384, 2013. 2
- [46] Haiyan Wang, Jiahao Pang, Muhammad A Lodhi, Yingli Tian, and Dong Tian. FESTA: Flow Estimation via Spatial-Temporal Attention for Scene Point Clouds. In *CVPR*, pages 14173–14182, 2021. 2
- [47] Zirui Wang, Shuda Li, Henry Howard-Jenkins, Victor Prisacariu, and Min Chen. Flownet3d++: Geometric losses for deep scene flow estimation. In *WACV*, pages 91–98, 2020. 2
- [48] Andreas Wedel, Clemens Rabe, Tobi Vaudrey, Thomas Brox, Uwe Franke, and Daniel Cremers. Efficient dense scene flow from sparse or dense stereo data. In *ECCV*, pages 739–751, 2008. 2
- [49] Yi Wei, Ziyi Wang, Yongming Rao, Jiwen Lu, and Jie Zhou. PV-RAFT: point-voxel correlation fields for scene flow estimation of point clouds. In *CVPR*, pages 6954–6963, 2021. 1, 2, 7
- [50] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. In *IROS*, pages 10359–10366, 2020. 5
- [51] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *CVPR*, pages 9621–9630, 2019. 1, 2
- [52] Wenxuan Wu, Zhi Yuan Wang, Zhuwen Li, Wei Liu, and Li Fuxin. PointPWC-Net: Cost Volume on Point Clouds for (Self-) Supervised Scene Flow Estimation. In *ECCV*, pages 88–107, 2020. 1, 2, 3, 6
- [53] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, pages 1983–1992, 2018. 2
- [54] Yin Zhou and Oncel Tuzel. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. In *CVPR*, pages 4490–4499, 2018. 1
- [55] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. DF-Net: Unsupervised Joint Learning of Depth and Flow using Cross-Task Consistency. In *ECCV*, pages 36–53, 2018. 2