

Fast Neural Scene Flow

Xueqian Li^{*1} Jianqiao Zheng¹ Francesco Ferroni^{*2} Jhony Kaesemeyer Pontes^{*3} Simon Lucey^{*1}
¹The University of Adelaide ²NVIDIA ³Latitude AI

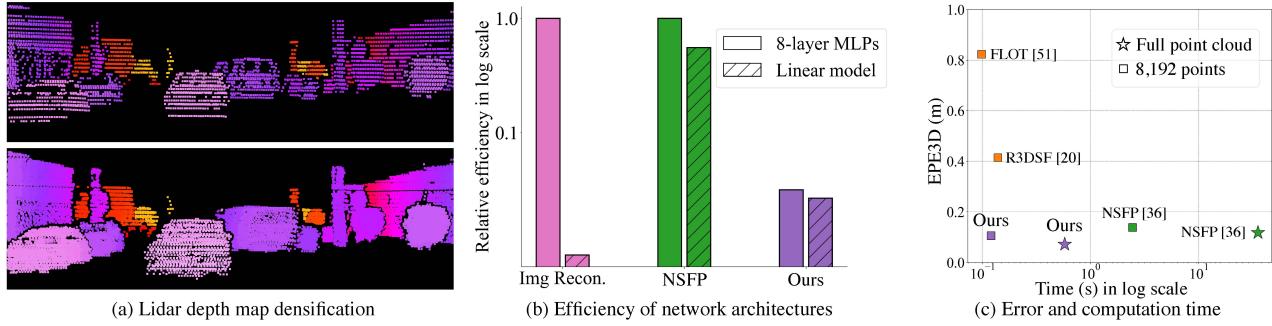


Figure 1. Scene flow is an important problem as it provides low-level motion cues for many downstream tasks. State-of-the-art learning methods are usually fast and can achieve impressive performance on in-domain data, but usually fail to generalize to out-of-the-distribution (OOD) data or handle dense point clouds. In this paper, we focus on a runtime optimization-based neural scene flow pipeline. In (a) one can see its application in the densification of lidar. However, in (c) one sees that the major drawback is the extensive computation time. We identify that the common speedup strategy in network architectures for coordinate networks has little effect on scene flow acceleration [see green (b)] unlike image reconstruction [see pink (b)]. With the dominant computational burden stemming instead from the Chamfer loss function, we propose to use a distance transform-based loss function to accelerate [see purple (b)], which achieves up to $30\times$ speedup and on-par estimation performance compared to NSFP [see (c)]. When tested on 8k points, it is as efficient [see (c)] as leading learning methods, achieving real-time performance.

Abstract

Neural Scene Flow Prior (NSFP) is of significant interest to the vision community due to its inherent robustness to out-of-distribution (OOD) effects and its ability to deal with dense lidar points. The approach utilizes a coordinate neural network to estimate scene flow at runtime, without any training. However, it is up to 100 times slower than current state-of-the-art learning methods. In other applications such as image, video, and radiance function reconstruction innovations in speeding up the runtime performance of coordinate networks have centered upon architectural changes. In this paper, we demonstrate that scene flow is different—with the dominant computational bottleneck stemming from the loss function itself (i.e., Chamfer distance). Further, we rediscover the distance transform (DT) as an efficient, correspondence-free loss function that dramatically speeds up the runtime optimization. Our fast neural scene flow (FNSF) approach reports for the first time real-time performance comparable to learning methods,

without any training or OOD bias on two of the largest open autonomous driving (AV) lidar datasets Waymo Open [64] and Argoverse [9].

1. Introduction

Neural Scene Flow Prior (NSFP) [38] is considered the dominant method in open-world perception [48] and scene densification [38, 71] using lidar (see Fig. 1 (a)). NSFP achieves state-of-the-art scene flow estimates from dense lidar point clouds (up to 150k+ points) and works on a variety of sensor setups with little to no refinement or adaptation. Unlike supervised or unsupervised learning-based methods, NSFP does not require learning from large offline datasets and has no limits on point density (<8k for most learning methods). Instead, it leverages the architecture of a neural network to implicitly regularize the flow estimate and employs a runtime optimization that can easily scale to large out-of-distribution (OOD) scenes, which is a challenge for current learning-based methods [16, 32, 38, 48, 52].

A fundamental drawback, however, to NSFP is the speed of its runtime optimization which is in some instances of orders of magnitude slower than its learning counterparts

^{*}Part of the work was done while at Argo AI. Corresponding e-mail: xueqian.li@adelaide.edu.au. Code available at <https://github.com/Lilac-Lee/FastNSF.git>

(see Fig. 1). As a result, NSFP has widely been used offline for (i) providing scene flow supervision for efficient learning methods and (ii) as a pre-processing step for training open-world perception systems [48]. However, the considerable computational cost of NSFP limits its current applications only to these offline tasks.

A central narrative of our approach is that the dominant computational burden in runtime scene flow optimization (NSFP) is not the network architecture, but the loss function—specifically Chamfer distance (CD) [18]. This differs considerably from other applications of coordinate networks throughout vision and learning such as neural radiance fields (NeRF [45]) and high-fidelity image reconstruction ([63]) which have gained significant speedups through architectural innovations [85]. A visual depiction of this discrepancy can be found in Fig. 1 (b).

Key to our approach is the use of correspondence-free loss function—distance transform (DT) [6, 14, 60] as a proxy for the computationally expensive CD loss. Even though DT has been extensively studied by the graphics and vision community over a few decades, its application as an efficient loss function in deep geometry has largely been overlooked up until this point. We believe that the inherent efficiencies of the DT are especially pertinent for runtime network training such as in NSFP. Our approach shares similarities with Plenoxels [83]—a recent approach for efficient radiance field estimation using coordinate networks—as we trade memory consumption for computation time, allowing for significant speedups during runtime, which provides an alternative solution when exploring more efficient loss functions for dense scene flow estimation. We differ from Plenoxels, however, in that our memory consumption stems from our proposed loss function, not the neural architecture itself.

In this paper, we present for the first time an approximately real-time (for 8k points) runtime optimization method as computationally efficient as leading learning methods whilst preserving the scalability to dense point clouds and state-of-the-art performance on OOD scenes like NSFP. We compare the performance and the computation time of our approach on two of the largest open lidar AV datasets available: Waymo Open [64] and Argoverse [9]. Our fast neural scene flow achieves up to ~ 30 times speedup than NSFP [38] (our faster implementation) and of comparable speed to leading learning-based methods (see Fig. 1 (c)) with the same number of points (8,192). It opens up the possibility of employing a fast, robust, and generalizable approach for dense scene flow, which is not prone to OOD effects in real-time vision and robotic applications.

2. Related work

Scene flow estimation. Scene flow denotes the motion field in 3D space [69] uplifted from 2D optical flow. To reconstruct 3D flow, traditional image-based methods [2, 25,

26, 28, 29, 37, 54] formulate an optimization problem utilizing RGB or depth information, and learning-based RGB/RGB-D methods [7, 30, 31, 58, 62, 66, 80] rely on single/multiple image features which encode with a large amount of data supervisions. On the other hand, to estimate scene flow directly from 3D, traditional point cloud-based methods [1, 12, 51] solve for a non-rigid registration problem, while recent work prefers full-supervised learning [24, 35, 39, 40, 53, 73, 75, 77] that uses point-based features or self-supervised learning [3, 22, 46, 67, 77] that employs a self-supervised loss. Recent non-learning-based methods [38, 52] draw our attention back to runtime optimization that easily scales to large data. Graph prior [52] explicitly builds a graph on the point cloud and uses a graph Laplacian regularizer. While neural scene flow prior [38] uses the network as an implicit regularizer to smooth motions. In this paper, we explore point cloud-based scene flow using runtime optimization.

Accelerating coordinate networks. There exists a line of work [10, 21, 27, 47, 57, 83, 84] that focuses on accelerating coordinate networks by trading slow, memory efficient, deep network architectures for fast, memory hungry, shallow architectures. Most of these innovations have been applied to the problem of neural radiance fields most notably Plenoxels [83] and TensorRF [10]. Recently, this trend was generalized for arbitrary signals through the introduction of complex positional encoding [85] with shallow linear networks. In this paper, we claim that these architectural innovations have little utility in speeding up neural scene flow without first addressing the computational cost of the Chamfer loss it uses.

Distance transform. DT [4, 6, 14, 42, 60] has played an important role in image processing, especially binary image analysis [49, 68]. Further applications are also found in medical image segmentation [13, 34, 61, 72, 74], robotics motion planning [56, 78], geometric representation [8, 11, 50], and accelerated point cloud registration [20, 81]. Among them, various distance measures have been used, such as city block, chessboard, and Euclidean distance [14, 79]. Naturally, Euclidean distance is preferred in computing point distance but it is also the most difficult metric to compute due to the temporal complexity [23]. Many work attempts to speedup Euclidean DT computation including raster-scan-based algorithms [6, 19, 36, 42, 55], fast marching-based algorithms [17, 41, 70], etc., and has achieved linear time computation. In this paper, we investigate the raster-scan-based algorithm for the 3D point cloud.

3. Approach

3.1. Background

Scene flow optimization. Suppose we have a moving sensor (*e.g.*, lidar mounted on a car, depth camera tied to a robot, *etc.*) collecting point cloud in a dynamic environment.

At time $t-1$, a point cloud \mathcal{S}_1 (source) of the scene is sampled. Then given the movements of the sensor and objects in certain directions, another point cloud \mathcal{S}_2 (target) is sampled at time t . In order to find out all the motions in the environment, we model the translation of each point cloud $\mathbf{p} \in \mathcal{S}_1$ from time $t-1$ to time t as a flow vector $\mathbf{f} \in \mathbb{R}^3$, where $\mathbf{p}' = \mathbf{p} + \mathbf{f}$. The translational vector set of all 3D points in \mathcal{S}_1 is defined as the scene flow $\mathcal{F} = \{\mathbf{f}_i\}_{i=1}^{|\mathcal{S}_1|}$.

Therefore, the optimization of the scene flow is to minimize the point distance between the source \mathcal{S}_1 and the target \mathcal{S}_2 . Usually, a regularization C , such as a Laplacian regularizer, is needed due to the highly unconstrained non-rigid flows. The overall optimization becomes

$$\mathcal{F}^* = \arg \min_{\mathcal{F}} \sum_{\mathbf{p} \in \mathcal{S}_1} D(\mathbf{p} + \mathbf{f}, \mathcal{S}_2) + \lambda C, \quad (1)$$

where D is a point distance function, λ is a coefficient of the regularizer C .

Neural scene flow prior. NSFP uses traditional runtime optimization to optimize a neural network. Different from learning-based methods, NSFP does not rely on any prior knowledge of large-scale datasets. And different from traditional scene flow optimization with an explicit regularizer that is mentioned above, neural scene flow prior optimizes parameters of a network which implicitly imposes a regularization by its structure:

$$\Theta^* = \arg \min_{\Theta} \sum_{\mathbf{p} \in \mathcal{S}_1} D(\mathbf{p} + g(\mathbf{p}; \Theta), \mathcal{S}_2). \quad (2)$$

where Θ is a parameter set of network g to be optimized. \mathbf{p} is the input source point, and the flow $\mathbf{f} = g(\mathbf{p}; \Theta)$ is the output of the network. The optimization converges at $\mathbf{f}^* = g(\mathbf{p}; \Theta^*)$. The network g here is chosen to be a commonly used RELU-MLP.

Since the points in source \mathcal{S}_1 and target \mathcal{S}_2 are not in correspondence, nor having the same number of points, *i.e.*, $|\mathcal{S}_1| \neq |\mathcal{S}_2|$, we use a distance function that handles these problems as

$$D(\mathbf{p}, \mathcal{S}) = \min_{\mathbf{x} \in \mathcal{S}} \|\mathbf{p} - \mathbf{x}\|_2^2. \quad (3)$$

Practically, a bidirectional point distance is found, yielding the above equation equivalent to the Chamfer loss [18].

Chamfer distance. In point cloud processing, Chamfer distance (CD) [18] is an important loss function and metric for computing the point distance of two point clouds that do not necessarily have points in correspondence. Chamfer distance loss computes the point distance of both source-to-target and target-to-source directions. In detail, the CD loss

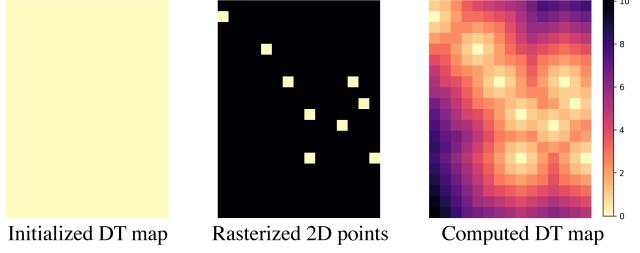


Figure 2. An example of how to build a DT map of 2D points. The initialized DT map is set to all zeros. The rasterized 2D point image is set to zero if a point is presented in the grid, and set to one if the grid contains no points. After computation, the final DT map is shown in the right figure.

can be written as

$$\begin{aligned} CD(\mathcal{S}_1, \mathcal{S}_2) &= \sum_{\mathbf{p} \in \mathcal{S}_2} D(\mathbf{p}, \mathcal{S}_1) + \sum_{\mathbf{q} \in \mathcal{S}_1} D(\mathbf{q}, \mathcal{S}_2) \\ &= \sum_{\mathbf{p} \in \mathcal{S}_2} \min_{\mathbf{x} \in \mathcal{S}_1} \|\mathbf{p} - \mathbf{x}\|_2^2 + \sum_{\mathbf{q} \in \mathcal{S}_1} \min_{\mathbf{y} \in \mathcal{S}_2} \|\mathbf{q} - \mathbf{y}\|_2^2. \end{aligned} \quad (4)$$

To compute the point distance, correspondences from source-to-target and target-to-source are searched among the nearest point neighbors. However, the exhaustive point correspondence search is extremely slow, especially when dealing with dense point clouds that contain more than 10k points (*e.g.*, in Waymo Open and Argoverse datasets, the number of points can be up to 150k+).

3.2. Correspondence-free point distance transform

DT is widely used in 2D image processing, such as segmentation, boundary detection, pattern matching, skeletonization, *etc.* However, general usage in irregular and unordered 3D point cloud tasks is not broadly discussed. Given that the image grid is regular and ordered, a DT map is easily obtained by computing the minimum distance of each sampled point $\mathbf{x} \in \mathcal{S}$ to the vertex \mathbf{q} of a DT map \mathcal{G} as

$$DT(\mathbf{q}) = \min_{\mathbf{x} \in \mathcal{S}, \mathbf{q} \in \mathcal{G}} D(\mathbf{x}, \mathbf{q}). \quad (5)$$

We extend Eq. (5) to fit in the scene flow task such that D refers to Euclidean distance, \mathbf{x} denotes the target point, and \mathbf{q} is the regularly spaced point in a voxel (3D).

Approximation of DT map. However, with a large number of points in a point cloud and the grid map, directly computing Eq. (5) builds high dimensional matrices which will lead to large memory occupation. Even when we presume that each axis of the grid/voxel is separable, and the distance per axis is pre-computed, the memory consumption is still huge that cannot be processed on a

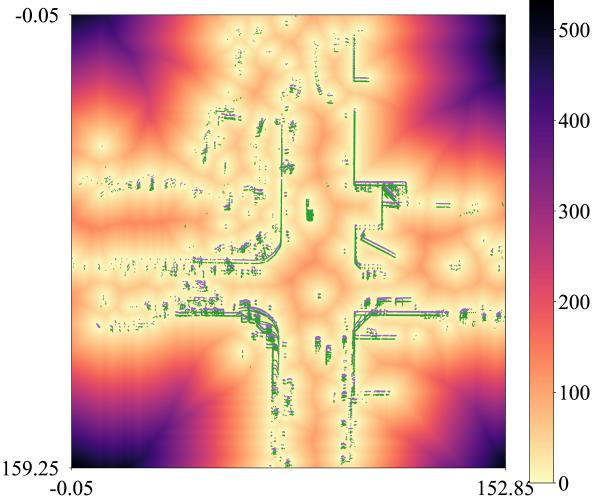


Figure 3. Distance transform map of a target point cloud in Argoverse dataset. Purple points are the target, and green points are the source. It is important to note that the DT map is constructed from the target points, and the source points are only shown for reference purposes. The colormap denotes the DT values (in meters), with light yellow denoting a smaller distance and dark purple indicating a larger distance.

single GPU. One strategy is to iteratively aggregate small matrix multiplications while at a cost of extensive time consumption. Another strategy is to use k-d tree query to compute the Euclidean distance. Building a k-d tree is one-time computing, but querying from a large point set to an even larger k-d tree is expensive.

Instead, we use a fast distance approximation by rasterizing the irregular points onto the nearest regular cuboid/grid. Specifically, similar to classical efficient binary image-based DT methods [13, 19], we construct two 3D binary images for target points and an initialized DT map. An example of 2D DT map is shown in Fig. 2. Specifically, the DT map is set to be all zeros, and the target binary image pixel is set to one when the grid contains no target points in it. Then a two-pass raster scan-based algorithm [19] is applied to each axis to compute the final DT, which means a 3D image needs six propagation passes (left to right, right to left, front to back, back to front, top to bottom, bottom to top). Although we rasterize target points using a binary 3D image approximation, we still compute the exact Euclidean distance between these two binary images instead of an approximated point distance. We empirically find that such discretization does not hurt performance. One reason is the local rigidity of the scene flow benefits from reasonably small but not necessarily infinitesimal DT grids/voxels.

A visualization of a 2D bird’s eye view (BEV) DT map is presented in Fig. 3, where we see that when the point density is large, the distance value is relatively small, and the distance value becomes extremely large when no points exist. By choosing an appropriate grid size, the

pattern of DT is distinct among different points while maintaining a relatively smooth structure in the neighboring area. Evidently, the correspondence-free distance transform can act as an effective and efficient surrogate for point correspondence-based Chamfer loss.

DT query and loss function. Once a DT map is built for the target point cloud, we can look up the pre-computed DT map to get the distance of the nearest source point in an extremely fast fashion. The loss function becomes the queried Euclidean distance between the target and the nearest deformed source points:

$$\mathcal{L} = \min_{\mathbf{y} \in \mathcal{S}', \mathbf{q} \in \mathcal{G}} \|\mathbf{y} - \mathbf{q}\|_2, \quad (6)$$

where \mathcal{S}' is the deformed source point cloud. Unlike CD loss, no point correspondence search is required in Eq. (6), making the DT query exceptionally fast. Further ablation studies can be found in Sec. 4.3.

4. Experiments

Datasets We are primarily interested in scene flow methods that are well-suited for large-scale, realistic, lidar-based OOD scenes, which are commonly encountered in autonomous driving (AV) applications. To this end, we focus on two AV datasets: Waymo Open [64] and Argoverse [9], which contain numerous challenging dynamic scenes. Unfortunately, no ground truth annotations were provided for the open-world dataset. We pre-processed the Argoverse and Waymo Open pseudo ground truth scene flow datasets following [38, 52] and [32, 82] respectively.

Metrics We follow scene flow metrics used in [38, 39, 46, 52, 77] to evaluate the performance. We also include the computation time breakdown in the table. These metrics are:

1) 3D end-point error $\mathcal{E}(m)$ that measures the mean absolute point distance between estimated and target points;

2) strict accuracy $Acc_5(\%)$, which is the accuracy of the estimated flow that satisfies the absolute point error $\mathcal{E} < 0.05m$ or the relative point error $\mathcal{E}' < 5\%$;

3) relaxed accuracy $Acc_{10}(\%)$, which is the accuracy of the estimated flow that satisfies the absolute point error $\mathcal{E} < 0.1m$ or the relative point error $\mathcal{E}' < 10\%$;

4) angle error $\theta_\epsilon(rad)$ that measures the mean angle error (in radian) of the estimated and the pseudo ground truth translational vectors.

5) computation time $t(ms)$ breaks down to four parts.

Pre-compute includes data loading and building a DT map. **Corr. / DT query** counts the time needed to search point correspondences or query point distance within a DT map. We also include computation time for **Network** forward and backward propagation. Finally, **Total** time in seconds / milliseconds is measured.

Implementation details We provide the details of implementation for each algorithm we compared. Further information can be found in the supplementary materials.

1) NSFP [38]. The original implementation is extremely slow. Based on the official code released by the authors while keeping the same parameter settings specified in the original paper, we implemented a faster version (NSFP) for a fair comparison. Note that to reflect the independence of each pair of point clouds when estimating scene flow, we randomly initialized the network before each optimization.

2) Baseline. Note that although the backward flow enforces a cycle consistency between the deformed point cloud and the original point cloud, we empirically found that removing cycle consistency does not hurt the overall performance when dealing with dense lidar point clouds, but improves the computational efficiency. Here we removed the backward flow in the original NSFP and implemented a **NSFP (baseline)** version as our baseline for a fair comparison—all computation times were compared with this model. We further modified the baseline model using a linear model with complex positional encodings (**NSFP (baseline, linear)**) to demonstrate the effect of network architecture changes. Detailed explanations can be found in Sec. 4.1 and the supplementary material.

3) Ours. We implemented a DT-based neural scene flow method with 8-layer ReLU-MLPs (**Ours**). We chose the grid cell size of the DT map to be 0.1 meters. Ablation studies on the choice of grid cell size can be found in Sec. 4.3. We further modified our method using a linear model with complex positional encodings (**Ours (linear)**) to demonstrate the effect of network architecture changes.

4) FlowStep3D [35] and FLOT [53] are fully supervised methods trained on synthetic FlyingThings3D [43] datasets. **PointPWC-Net [77]** can be used as a self-supervised method. We used the official code released by the authors and directly tested the pretrained model (pretrained on FlyingThings3D) on our datasets. However, as full/self-learning-based methods, they performed poorly on OOD datasets, which is also observed in [16, 32, 38, 48, 52]. Here we only chose the method with the best performance and the lowest computation time—FLOT—for comparison in the main table. The comparison of other learning-based methods is included in the supplementary material. **R3DSF [22]** is a weakly supervised method with no direct supervision of ground truth dynamic flows. We tested the method using the pretrained model (on KITTI [44]) provided by the authors.

All models were implemented using CUDA 11.6-supported PyTorch. All experiments were run on a computer with a single NVIDIA RTX 3090Ti GPU and a 24 AMD Ryzen 9 5900X 12-Core CPU @ 4.95GHz.

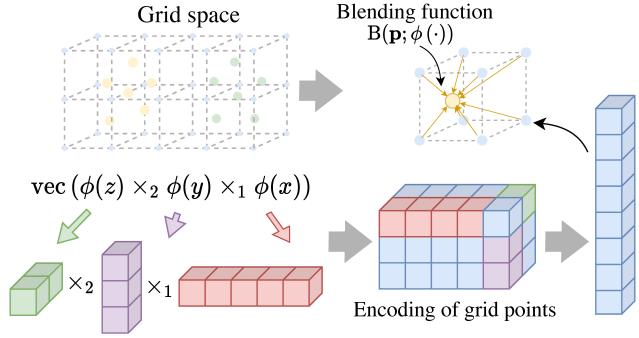


Figure 4. Illustration of 3D complex PE in scene flow problem. The 3D space of source points (yellow) and target points (green) is divided into small cubes. The number of cubes on each axis differs but the edge length is the same. Encoding of a point in a certain grid is the blending/interpolation of encodings of all 8 grid vertices. These grid points (blue) are encoded by complex PE along each axis. Note the vec and outer product (\times) notation is for better visual understanding and it is equivalent to the Kronecker product.

4.1. Speedup in network architecture

Plenoxels [83] is a method that accelerates NeRF [45] optimization by replacing deep neural networks with voxel-based spherical harmonics representations, leading to efficient volume rendering. Similar network speedup approaches can also be used in neural scene flow optimization. We incorporate recent innovations in positional encodings, specifically complex positional encoding (complex PE) [85] to represent high-frequency signals and enable a linear reconstruction model which is similar in spirit to [83].

Complex positional encodings Positional encodings (PEs) are encodings for input positions—*e.g.*, 2D image grids, 3D voxel grids, *etc.*—that are usually used in coordinate networks. PEs have been shown to improve the performance and convergence speed of coordinate networks [65]. Simple PE is a simple concatenation of the encoding in each input dimension, while a complex PE is a more complicated encoding that computes the Kronecker product of the per-dimension encoding. An illustration of complex PE of scene flow is shown in Fig. 4.

With a complex PE and a linear model parameterized by $\mathbf{W} \in \mathbb{R}^{W_x W_y W_z \times 3}$, the scene flow can be represented as

$$\mathbf{f} \approx \mathbf{B}(\mathbf{p}; \phi) \text{vec} (\phi(\mathbf{x}) \mathbf{W} \phi(\mathbf{z})^T \phi(\mathbf{y})^T), \quad (7)$$

where $\mathbf{B}(\cdot)$ is the blending function, $\phi(\cdot)$ is the encoder, $\phi(\mathbf{x}) \mathbf{W} \phi(\mathbf{z})^T \phi(\mathbf{y})^T$ is a simple notation for n -mode multiplication, which is equivalent to Kronecker product. The blending function interpolates the 3D regular grid encodings to handle the irregular and unordered nature of 3D point clouds. Additional details of complex PE can be found in the supplementary materials.

Table 1. Computation time and performance on Waymo Open Scene Flow dataset. We generated 202 testing scene flow examples where each point cloud has 8k-144k points. The upper tabular between **blue bars** are experiments with the full point cloud, and the lower tabular between **orange bars** are experiments with only 8,192 points. All query time is listed as time per optimization step [total query time]. The speedups are marked with **green \times** numbers, and the slow-downs are marked with **red \times** . The computation time of using full point cloud is compared to values in the **red box**, and the computation time of using 8,192 points is compared to values in **green box**. Bold texts are the best performance, and underlined texts denote the second-best performance. \uparrow means larger values are better while \downarrow means smaller values are better. Corr. / DT query denotes correspondence search or DT query. **Note:** performance of learning methods PointPWC-Net [77], FlowStep3D [35], PV-RAFT [76] can be found in the supplementary material.

Method	\mathcal{E}	Acc_5	Acc_{10}	θ_e	$t (ms) \downarrow$			
	(m) \downarrow	(%) \uparrow	(%) \uparrow	(rad) \downarrow	Pre-compute	Corr. / DT query	Network	Total
NSFP [38]	0.100	<u>76.62</u>	88.56	<u>0.286</u>	—	114 [30972]	4.62 [1361]	35.51 s
NSFP (baseline)	0.118	74.16	86.70	0.300	—	43.1 [15036]	2.38 [904]	18.39 s
NSFP (baseline, linear)	<u>0.096</u>	70.78	86.31	0.310	9.48	40.38 [8037] 1.1\times	1.53 [319] 1.6\times	10.20 s 1.80\times
Ours	0.072	84.73	92.24	0.280	40.85	0.25 [13] 1.72\times	2.89 [149] 1.2\times	0.58 s 31.7\times
Ours (linear)	0.109	71.27	85.80	0.321	44.88	0.23 [19] 1.87\times	1.62 [138] 1.5\times	0.49 s 37.5\times
FLOT [53]	0.702	2.46	11.30	0.808	—	—	—	99 ms
R3DSF [22]	0.414	35.47	44.96	0.527	—	—	—	140 ms
NSFP [38] (8,192 pts)	<u>0.138</u>	<u>53.62</u>	<u>78.57</u>	<u>0.339</u>	—	5.42 [1285] 21\times	4.44 [1051] 1.1\times	2459 ms 17.6\times
Ours (8,192 pts)	0.106	77.53	88.99	0.329	35.22	0.23 [6.5] 496\times	2.60 [76] 1.8\times	121 ms 1.16\times

Table 2. Computation time and performance on Argoverse Scene Flow dataset. Argoverse has 212 testing scene flow examples where each point cloud has 30k-70k points. Notations are the same as in the table above.

Method	\mathcal{E}	Acc_5	Acc_{10}	θ_e	$t (ms) \downarrow$			
	(m) \downarrow	(%) \uparrow	(%) \uparrow	(rad) \downarrow	Pre-compute	Corr. / DT query	Network	Total
NSFP [38]	0.069	<u>71.56</u>	87.80	0.235	—	47 [14310]	4.66 [1507]	18.08 s
NSFP (baseline)	0.078	69.46	86.22	<u>0.253</u>	—	17 [5901]	2.31 [848]	8.38 s
NSFP (baseline, linear)	0.097	67.03	83.20	0.314	9.19	14.7 [2786] 1.1\times	1.51 [297] 1.5\times	3.55 s 2.36\times
Ours	<u>0.071</u>	80.05	90.71	0.289	43.61	0.24 [14] 71\times	2.57 [149] 1.1\times	0.51 s 16.4\times
Ours (linear)	0.106	65.00	82.85	0.319	48.59	0.23 [20] 74\times	1.69 [149] 1.4\times	0.43 s 19.5\times
FLOT [53]	0.821	2.00	8.84	0.967	—	—	—	88 ms
R3DSF [22]	0.417	32.52	42.52	0.551	—	—	—	113 ms
NSFP [38] (8,192 pts)	0.113	<u>46.32</u>	<u>72.68</u>	0.347	—	5.40 [1500] 8.7\times	4.42 [1233] 1.1\times	2864 ms 25.6\times
Ours (8,192 pts)	0.118	69.93	83.55	0.352	41.57	0.22 [6.33] 214\times	2.51 [72.69] 1.9\times	<u>124 ms</u> 1.10\times

Network speedup comparison We would like to point out that although various strategies have been proposed to accelerate network architectures, they do not lead to substantial speedup in neural scene flow estimation. To demonstrate this, we compare the relative efficiency of using 8-layer Relu-MLPs and complex PE with a linear model in both 2D image reconstruction and 3D scene flow tasks, shown in Fig. 1 (b). In detail, following [85], we used a size of 256×256 image dataset for image reconstruction. For a fair comparison, the relative efficiency is obtained by normalizing the computation time of two tasks. The computation time of single image reconstruction is 17.63 s and 0.15 s for deep network and linear network respectively, leading to $118\times$ speedup. However, the linear model only results in $\sim 2\times$ and $\sim 1.2\times$ speedup for NSFP and our method respectively. Our results show that complex PE-based speedup achieves significant acceleration in 2D image reconstruction compared to deep neural networks, but the

benefits of such speedup are more limited when estimating neural scene flow. More detailed comparisons are in Sec. 4.2.

4.2. Comparison of performance

We show the performance of our method with different variants compared to NSFP, FLOT, and R3DSF on Waymo Open (Tab. 1) and Argoverse (Tab. 2) scene flow datasets. We denote results on Waymo Open (xx) and Argoverse (yy) as xx/yy. Visual results and applications of densification are shown in Fig. 5 and supplementary materials.

Dense scene flow estimation. For dense point clouds, the baseline NSFP took 18.39/8.38 s to converge. When replacing the 8-layer MLPs with a linear network and applying complex PE, we observed a speedup of only 1.80/2.36 \times in total, primarily due to the network propagation speedup. When we replaced naive CD loss with the DT loss, a significant 31.7/16.4 \times speedup was achieved while

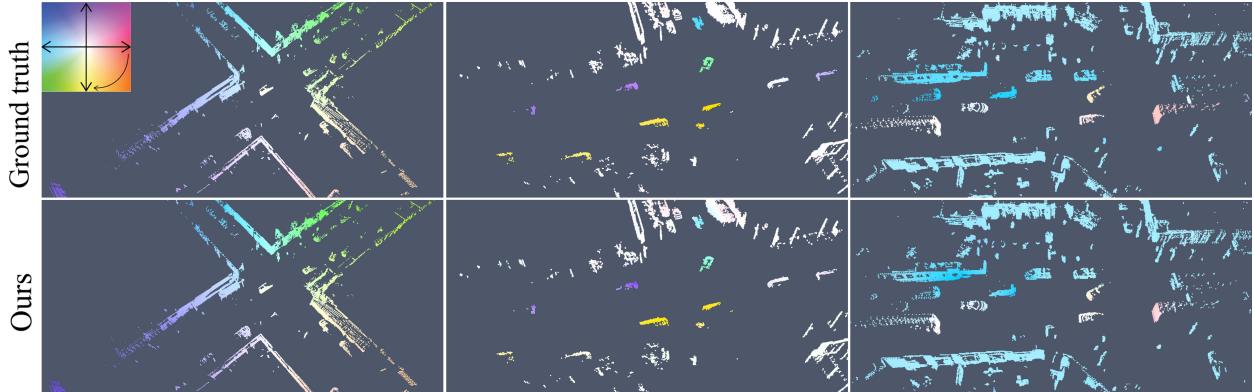


Figure 5. Visual examples of the scene flow prediction using our method on the Argoverse dataset. We show 3 different autonomous driving scenarios where the autonomous vehicle (AV) was making a right turn (left), the AV stopped for crossing traffic (middle), and the AV was driving in the city (right). We found that the scene flow predicted by our method is close to the ground truth flows. The upper left corner color wheel indicates the flow magnitude (color intensity) and the flow direction (angle). For example, the yellow vehicles in the middle figure are heading south with a relatively large speed wrt the AV.

maintaining comparable high accuracy. However, replacing a deep network with a linear model when using DT loss only resulted in an additional $1.2/1.2 \times$ speedup at the cost of a decrease in flow accuracy. Specifically, the total computation time of our proposed method with DT loss is 0.58/0.51 s, and with DT loss and a linear model is 0.49/0.43 s. However, with a marginal decrease in computation time, we observed a relatively large drop in performance such that the strict accuracy of our method with DT is 84.73/80.05% and 71.27/65.00% with DT and a linear model. These results strongly support our argument in the previous section, which is the general strategy of simplifying or replacing network architectures to accelerate coordinate networks is not particularly effective when optimizing scene flow through coordinate networks.

Overall, the performance of our method using DT loss has on-par performance compared to NSFP while being orders of magnitude faster. We have noticed that there is an “improved accuracy” of our method compared to NSFP in Tab. 1, Tab. 2 based on these facts: 1. The performance of our method and NSFP is similar—our method achieves slightly higher performance on some metrics (*e.g.*, accuracy), and NSFP achieves slightly higher performance on other metrics (*e.g.*, angular error). These results indicate that distance transform (DT) is an effective and efficient surrogate for the Chamfer distance (CD) in the scene flow problem. 2. Compared to the “exact” point distance (CD), DT queries the distance based on a voxelized DT map, which naturally smooths the flow estimation. In real-world applications, we believe using a DT-accelerated deep network model will achieve both high-fidelity accuracy and efficient computation.

Computation time breakdown. A detailed time breakdown is provided in both tables, including pre-computation, correspondence search / DT distance query, and network forward / backward propagation. The naive

Chamfer distance requires per-point correspondence search, which can be extremely slow (43.1/17 ms per iteration), especially when the point cloud is denser. For instance, correspondence search is much slower in the Waymo Open dataset than in the Argoverse dataset since Waymo Open has an average of 90k points while Argoverse has an average of 50k points in a single point cloud. DT query is a pre-defined table lookup that is exceptionally faster than the naive point correspondence search, achieving a remarkable 172/71 \times speedup per optimization step—the main contributor to the overall efficiency of our method. Note that the pre-computing of the DT map (\sim 40 ms) is acceptable given that it is one-time computing and the total computation time is in the range of hundreds of milliseconds. Additional complex PE and a linear model only provide a modest speedup of 1.6/1.5 \times in the network propagation step. Further optimization of the pre-computation and faster implementation can be achieved with full CUDA support. Note that although in NSFP, the CD loss was implemented using PyTorch3D with CUDA acceleration, it still requires significant computation time, indicating its inherent limitations that cannot be easily overcome through engineering techniques alone.

OOD generalizability. We further extend our model using fewer points (8,192 points) to accommodate for a fair comparison against learning-based methods [22, 35, 53, 76, 77]. For these learning models, a fixed number of points is required and they only operate on fewer points, such as 2,048 [39, 53] or 8,192 points [22, 35, 53, 77]. Also, these learning methods need to crop the point cloud to a small range. Therefore, learning models cannot be easily adapted to large-scale dense point clouds. Jund *et al.* explored to inference dense point cloud [33], but they did not directly process the full point cloud as the input. A simple solution to address this challenge is to iteratively predict scene flow for

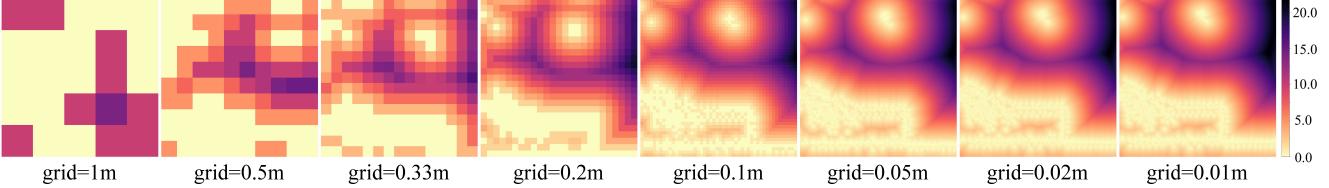


Figure 6. 2D DT map with different grid sizes. A larger grid size results in extremely coarse DT, as the grid cell size decreases, the DT accuracy increases. Here we zoom in on the original DT map in Fig. 3 to show the detailed distance values.

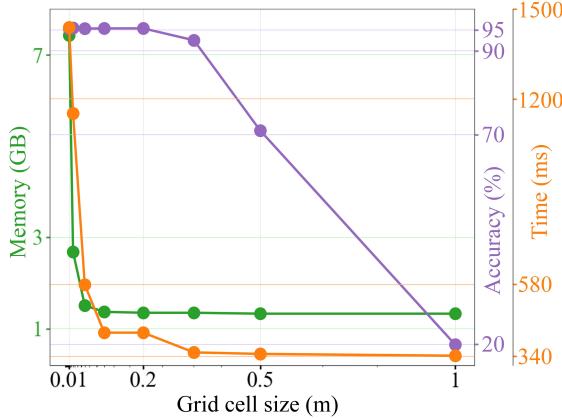


Figure 7. Performance of distance transform with different grid cell sizes on a 2D scene from Argoverse scene flow dataset. Generally, the larger the grid cell is, the lower the accuracy becomes. The computation time and memory consumption also decrease as the grid cell size increases (the number of grids decreases). The total grid of the DT map is of size 160×150.

small subsets of dense point clouds using learning methods. However, this approach requires intensive computation, and it might cause out-of-memory issues [33].

Fully supervised learning methods FlowStep3D, FLOT, PV-RAFT, and self-supervised method PointPWC-Net do not generalize to new autonomous driving datasets (see Tab. 1, Tab. 2) due to the domain gap between the training and testing datasets [16, 32, 38, 48, 52]. We provide additional performance in the supplementary material. Despite its poor performance on OOD data, FLOT achieves a competitive inference time, making it one of the fastest learning-based methods available. Note that since the weakly supervised learning method R3DSF takes supervision from object segmentation and AV ego motions and was trained on a lidar point cloud-based dataset, it has a relatively high accuracy compared to full/self-supervised learning methods, but its performance is still inferior to non-learning-based methods NSFP and our method. Such evidence strongly suggests that our method—a runtime optimization—is robust to OOD effects, and can be directly applied to applications where no training data is readily available.

Towards real-time computation. Our method attains real-time performance (121/124 ms) which is comparable to learning methods and maintains high accuracy, while

Table 3. Comparison of our method and NSFP++ on Argoverse scene flow dataset with ego-motion compensation.

Method	$\mathcal{E} \downarrow$ (m)	$Acc_5 \uparrow$ (%)	$Acc_{10} \uparrow$ (%)	$\theta_\epsilon \downarrow$ (rad)	$t \downarrow$ (ms)
Ours	0.411	34.94	46.82	0.731	335
NSFP++ [48]	0.295	<u>31.82</u>	62.61	0.343	16188
NSFP++ (DT)	0.272	30.26	60.25	0.365	2001

learning-based methods struggle to generalize on OOD data. It indicates the significant potential of applying robust and accurate runtime optimization-based methods in many vision-based applications.

Other learning methods. Learning-based scene flow methods are usually fast and achieve good accuracy when applied to in-domain small-scale data (training and testing are on the same dataset with a specific range: *e.g.*, KITTI data with point clouds within 35m of the scene center) with a limited number of points (usually 2,048 / 8,192 points). As discussed in [16, 32, 38, 48, 52], the domain gap is a significant challenge for learning methods—a specific dataset with specific configurations—*e.g.*, coordinate systems, viewing directions, *etc.*—that match the testing data needs to be used during training. However, we are interested in exploring runtime optimization-based methods that are robust to large-scale OOD data that can be employed in many real-world applications, such as autonomous driving scenarios, where no labels are readily available.

4.3. Ablation study of DT grid size

DT splits the space into small grid cells, while the grid size affects the accuracy, the computation time, and the memory accuracy. We provide a performance comparison using different grid cell sizes of DT in Fig. 7. A 2D visualization of DT with different grid sizes is also shown in Fig. 6. We can clearly see that a relatively small grid cell size ($grid \leq 0.1m$) is required to ensure the fidelity of the distance transform map. Note that when rasterized points are closer to the original irregular points, DT is a closer approximation to the exact Euclidean distance between two point clouds. Moreover, with a rasterized point-based DT strategy, the pre-computation of the map is no longer an overhead, and the grid cell size will largely affect the memory instead of the computation time—we trade memory consumption with computation time.

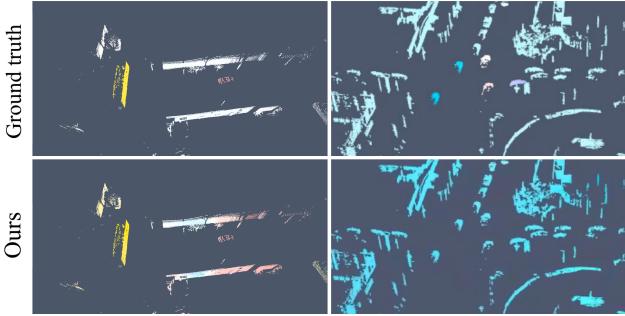


Figure 8. Failure cases on Argoverse dataset. The left figure shows some noisy motions were predicted in the background. The right figure shows our method failed to predict when the motion of some cars is relatively small to the AV.

4.4. Speedup in other methods

The proposed DT loss is a *general* loss that can be used in other 3D geometry tasks. Specifically, for optimization-based methods such as RSF [15], DT could replace the nearest neighbor distance loss to speed up the overall computation. For learning methods such as PointPWCNet [77], DT could replace Chamfer loss to speed up optimization during training, especially when dealing with a large number of points. For methods using cycle consistency, when replacing CD ($\mathcal{O}(N^2)$) with DT, the forward flow computation will be substantially speeded up (similar speedup to the paper). The backward flow computation will also be speeded up even though the DT map will be computed in every optimization step.

We provide an example of comparing the results of our method to NSFP++ [48] in Tab. 3. We used our version of NSFP++ due to no publicly available official implementation. Since NSFP++ requires ego-motion compensation, we also created an additional Argoverse scene flow dataset that removes the ego motions of autonomous vehicles. These ego motions are provided in the original Argoverse dataset [9].

Following [48], we define dynamic points as points that the norm between the deformed source (*i.e.*, source point cloud deformed by the ground truth flow) and the ego-motion compensated source (*i.e.*, the source point cloud is transformed by the ego-motion) exceeds a threshold of 0.05. All the metrics are computed only on dynamic points. During optimization, for our method, we removed the ego motion of the point cloud and estimated the scene flow using the full point cloud, and for NSFP++, we only used dynamic points as required.

In Tab. 3, we show that our method achieves worse performance than NSFP++ while being $\sim 50\times$ faster. We further replaced the Chamfer distance loss used in NSFP++ with our proposed distance transform (DT) loss, and we observed a $\sim 8\times$ speedup. The distance transform loss can be a robust and efficient surrogate to Chamfer distance loss in many deep geometry vision tasks.

4.5. Limitations

One drawback of our method is that creating a DT map using rasterization can lead to discretization errors, especially when the grid size of the DT map is large. To mitigate these, it is necessary to build a DT map with relatively fine-resolution grids. In this case, the memory consumption will increase especially when dealing with high-dimensional data, such as 3D point clouds. Further engineering efforts of pre-creating an efficient high-resolution DT map are required to maintain a reasonable memory cost. However, we empirically find that in the context of scene flow estimation—specifically for scene flow in autonomous driving scenarios—a relatively smooth representation is preferred for local rigidity assumptions. Nevertheless, the tradeoff between the grid resolution, the memory consumption, and the estimation accuracy should be carefully considered and chosen based on specific real-world applications.

Failure cases. We show two typical failure cases in Fig. 8. The first case (left column) is when dynamic points are much sparser than the background, our prediction can result in noisy non-rigid motions in the background. The second case (right column) is when the dynamic motion is relatively small, our model may fail to recognize the dynamic scene and only predict rigid motions.

5. Conclusion

In this work, we revisit the runtime optimization-based scene flow method NSFP and propose a method that is both efficient and generalizable to large-scale OOD data and dense lidar points. We identify that common strategies for speeding up network architectures do not yield significant time reductions—the major computation overhead is the Chamfer distance loss. Therefore, we propose to use an efficient correspondence-free distance transform loss as a robust surrogate. The rediscovery of DT in scene flow estimation opens up an innovative venue to leverage its efficiency and robustness for various deep geometry tasks. Compared to NSFP, our method maintains comparable accuracy but gains up to $\sim 30\times$ speedups. We report for the first time a real-time performance (~ 120 ms) with neural scene flow and runtime optimization when using fewer points (8,192). The efficient runtime optimization-based neural scene flow can be widely applied in lidar scenes to do point cloud densification, open-world object detection, and scene clustering, such as in autonomous driving scenarios, where no ground truth or training data are readily available.

Acknowledgement: We would like to thank Haosen Xing for the careful review of the manuscript and the help throughout the project. We thank Kavisha Vidanapathirana for the initial implementation of NSFP++ [48].

References

- [1] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid ICP algorithms for surface registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007. [2](#)
- [2] Tali Basha, Yael Moses, and Nahum Kiryati. Multi-view scene flow estimation: A view centered variational approach. *International Journal of Computer Vision (IJCV)*, 101(1):6–21, 2013. [2](#)
- [3] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13126–13136, 2021. [2](#)
- [4] Gunilla Borgefors. On digital distance transforms in three dimensions. *Computer vision and image understanding*, 64(3):368–376, 1996. [2](#)
- [5] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018. [15](#)
- [6] Heinz Breu, Joseph Gil, David Kirkpatrick, and Michael Werman. Linear time euclidean distance transform algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):529–533, 1995. [2](#)
- [7] Fabian Brickwedde, Steffen Abraham, and Rudolf Mester. Mono-SF: Multi-view geometry meets single-view depth for monocular scene flow estimation of dynamic traffic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2780–2790, 2019. [2](#)
- [8] Tony Chan and Wei Zhu. Level set based shape prior segmentation. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, volume 2, pages 1164–1170. IEEE, 2005. [2](#)
- [9] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3D tracking and forecasting with rich maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8748–8757, 2019. [1, 2, 4, 9, 16](#)
- [10] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. *arXiv preprint arXiv:2203.09517*, 2022. [2](#)
- [11] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5939–5948, 2019. [2](#)
- [12] Haili Chui and Anand Rangarajan. A new point matching algorithm for non-rigid registration. *Computer Vision and Image Understanding (CVIU)*, 89(2-3):114–141, 2003. [2](#)
- [13] Antonio Criminisi, Toby Sharp, and Andrew Blake. Geos: Geodesic image segmentation. In *European Conference on Computer Vision*, pages 99–112. Springer, 2008. [2, 4](#)
- [14] Per-Erik Danielsson. Euclidean distance mapping. *Computer Graphics and image processing*, 14(3):227–248, 1980. [2](#)
- [15] David Deng and Avideh Zakhor. Rsf: Optimizing rigid scene flow from 3d point clouds without labels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1277–1286, 2023. [9](#)
- [16] Guanting Dong, Yueyi Zhang, Hanlin Li, Xiaoyan Sun, and Zhiwei Xiong. Exploiting rigidity constraints for lidar scene flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12776–12785, 2022. [1, 5, 8](#)
- [17] Hinnik Eggers. Two fast euclidean distance transformations in z2based on sufficient propagation. *Computer Vision and Image Understanding*, 69(1):106–116, 1998. [2](#)
- [18] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3D object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 605–613, 2017. [2, 3](#)
- [19] Pedro F Felzenszwalb and Daniel P Huttenlocher. Distance transforms of sampled functions. *Theory of computing*, 8(1):415–428, 2012. [2, 4](#)
- [20] Andrew W Fitzgibbon. Robust registration of 2d and 3d point sets. *Image and vision computing*, 21(13-14):1145–1153, 2003. [2](#)
- [21] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14346–14355, 2021. [2](#)
- [22] Zan Gojcic, Or Litany, Andreas Wieser, Leonidas J Guibas, and Tolga Birdal. Weakly supervised learning of rigid 3d scene flow. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5692–5703, 2021. [2, 5, 6, 7, 15](#)
- [23] George J Grevera. Distance transform algorithms and their implementation and evaluation. In *Deformable Models*, pages 33–60. Springer, 2007. [2](#)
- [24] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3254–3263, 2019. [2](#)
- [25] Simon Hadfield and Richard Bowden. Kinecting the dots: Particle based scene flow from depth sensors. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 2290–2295. IEEE, 2011. [2](#)
- [26] Simon Hadfield and Richard Bowden. Scene particles: Unregularized particle-based scene flow estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(3):564–576, 2013. [2](#)
- [27] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5875–5884, 2021. [2](#)
- [28] Michael Hornacek, Andrew Fitzgibbon, and Carsten Rother. SphereFlow: 6 DoF scene flow from RGB-D pairs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3526–3533, 2014. [2](#)

- [29] Frédéric Huguet and Frédéric Devernay. A variational method for scene flow estimation from stereo sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–7. IEEE, 2007. 2
- [30] Junhua Hur and Stefan Roth. Self-supervised multi-frame monocular scene flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2684–2694, 2021. 2
- [31] Huazhu Jiang, Deqing Sun, Varun Jampani, Zhaoyang Lv, Erik Learned-Miller, and Jan Kautz. SENSE: A shared encoder network for scene-flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3195–3204, 2019. 2
- [32] Zhao Jin, Yinjie Lei, Naveed Akhtar, Haifeng Li, and Munawar Hayat. Deformation and correspondence aware unsupervised synthetic-to-real scene flow estimation for point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7233–7243, 2022. 1, 4, 5, 8
- [33] Philipp Jund, Chris Sweeney, Nichola Abdo, Zhifeng Chen, and Jonathon Shlens. Scalable scene flow from point clouds in the real world. *IEEE Robotics and Automation Letters*, 7(2):1589–1596, 2021. 7, 8
- [34] Davood Karimi and Septimiu E Salcudean. Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Transactions on medical imaging*, 39(2):499–513, 2019. 2
- [35] Yair Kittenplon, Yonina C Eldar, and Dan Raviv. FlowStep3D: Model unrolling for self-supervised scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 5, 6, 7, 15, 16
- [36] F Leymarie and Martin D Levine. Fast raster scan distance propagation on the discrete rectangular lattice. *CVGIP: Image Understanding*, 55(1):84–94, 1992. 2
- [37] Rui Li and Stan Sclaroff. Multi-scale 3D scene flow from binocular stereo sequences. *Computer Vision and Image Understanding (CVIU)*, 110(1):75–90, 2008. 2
- [38] Xueqian Li, Jhony Kaesemel Pontes, and Simon Lucey. Neural scene flow prior. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2, 4, 5, 6, 8, 15, 17
- [39] Xingyu Liu, Charles R Qi, and Leonidas J Guibas. FlowNet3D: Learning scene flow in 3D point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 529–537, 2019. 2, 4, 7
- [40] Xingyu Liu, Mengyuan Yan, and Jeannette Bohg. MeteorNet: Deep learning on dynamic 3d point cloud sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9246–9255, 2019. 2
- [41] Sean Mauch. A fast algorithm for computing the closest point and distance transform. *Go online to <http://www.acm.caltech.edu/seanm/software/cpt/cpt.pdf>*, 2000. 2
- [42] Calvin R Maurer, Rensheng Qi, and Vijay Raghavan. A linear time algorithm for computing exact euclidean distance transforms of binary images in arbitrary dimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):265–270, 2003. 2
- [43] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, 2016. 5
- [44] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3061–3070, 2015. 5, 16
- [45] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–421. Springer, 2020. 2, 5, 13
- [46] Himangi Mittal, Brian Okorn, and David Held. Just go with the flow: Self-supervised scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11177–11185, 2020. 2, 4
- [47] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. 2
- [48] Mahyar Najibi, Jingwei Ji, Yin Zhou, Charles R Qi, Xincheng Yan, Scott Ettlinger, and Dragomir Anguelov. Motion inspired unsupervised perception and prediction in autonomous driving. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pages 424–443. Springer, 2022. 1, 2, 5, 8, 9
- [49] C Wayne Niblack, Phillip B Gibbons, and David W Capson. Generating skeletons and centerlines from the distance transform. *CVGIP: Graphical Models and image processing*, 54(5):420–437, 1992. 2
- [50] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019. 2
- [51] Mark Pauly, Niloy J Mitra, Joachim Giesen, Markus H Gross, and Leonidas J Guibas. Example-based 3D scan completion. In *Symposium on Geometry Processing*, pages 23–32, 2005. 2
- [52] Jhony Kaesemel Pontes, James Hays, and Simon Lucey. Scene flow from point clouds with or without learning. In *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, 2020. 1, 2, 4, 5, 8, 15
- [53] Gilles Puy, Alexandre Boulch, and Renaud Marlet. FLOT: Scene Flow on Point Clouds Guided by Optimal Transport. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 5, 6, 7, 16
- [54] Julian Quiroga, Thomas Brox, Frédéric Devernay, and James Crowley. Dense semi-rigid scene flow estimation from rgbd images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 567–582. Springer, 2014. 2
- [55] Ingemar Ragnemalm. The euclidean distance transform in arbitrary dimensions. *Pattern Recognition Letters*, 14(11):883–888, 1993. 2

- [56] Nathan Ratliff, Matt Zucker, J Andrew Bagnell, and Siddhartha Srinivasa. Chomp: Gradient optimization techniques for efficient motion planning. In *2009 IEEE International Conference on Robotics and Automation*, pages 489–494. IEEE, 2009. 2
- [57] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14335–14345, 2021. 2
- [58] Rishav Rishav, Ramy Battrawy, René Schuster, Oliver Wasenmüller, and Didier Stricker. DeepLiDARFlow: A deep learning architecture for scene flow estimation using monocular camera and sparse LiDAR. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, pages 10460–10467. IEEE, 2020. 2
- [59] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. 14
- [60] Azriel Rosenfeld and John L Pfaltz. Sequential operations in digital picture processing. *Journal of the ACM (JACM)*, 13(4):471–494, 1966. 2
- [61] Mikael Rousson and Nikos Paragios. Shape priors for level set representations. In *European Conference on Computer Vision*, pages 78–92. Springer, 2002. 2
- [62] Lin Shao, Parth Shah, Vikrant Dwaracherla, and Jeannette Bohg. Motion-based object segmentation based on dense RGB-D scene flow. *IEEE Robotics and Automation Letters*, 3(4):3797–3804, 2018. 2
- [63] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Neural Information Processing Systems (NeurIPS)*, 33, 2020. 2
- [64] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 1, 2, 4, 16
- [65] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Neural Information Processing Systems (NeurIPS)*, 2020. 5, 13
- [66] Zachary Teed and Jia Deng. RAFT-3D: Scene flow using rigid-motion embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8375–8384, 2021. 2
- [67] Ivan Tishchenko, Sandro Lombardi, Martin Oswald, and Marc Pollefeys. Self-supervised learning of non-rigid residual flow and ego-motion. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2020. 2
- [68] Son Tran and Liwen Shih. Efficient 3d binary image skeletonization. In *2005 IEEE Computational Systems Bioinformatics Conference-Workshops (CSBW'05)*, pages 364–372. IEEE, 2005. 2
- [69] Sundar Vedula, Simon Baker, Peter Rander, Robert Collins, and Takeo Kanade. Three-dimensional scene flow. In *Proceedings of the International Conference on Computer Vision (ICCV)*, volume 2, pages 722–729. IEEE, 1999. 2
- [70] Ben J. H. Verwer, Piet W. Verbeek, and Simon T. Dekker. An efficient uniform cost algorithm applied to distance transforms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(4):425–429, 1989. 2
- [71] Chaoyang Wang, Xueqian Li, Jhony Kaesemel Pontes, and Simon Lucey. Neural prior for trajectory estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6532–6542, 2022. 1
- [72] Guotai Wang, Maria A Zuluaga, Wenqi Li, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Deepigeos: a deep interactive geodesic framework for medical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1559–1572, 2018. 2
- [73] Haiyan Wang, Jiahao Pang, Muhammad A Lodhi, Yingli Tian, and Dong Tian. Festa: Flow estimation via spatial-temporal attention for scene point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14173–14182, 2021. 2
- [74] Yan Wang, Xu Wei, Fengze Liu, Jieneng Chen, Yuyin Zhou, Wei Shen, Elliot K Fishman, and Alan L Yuille. Deep distance transform for tubular structure segmentation in ct scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3833–3842, 2020. 2
- [75] Zirui Wang, Shuda Li, Henry Howard-Jenkins, Victor Prisacariu, and Min Chen. FlowNet3D++: Geometric losses for deep scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 91–98, 2020. 2
- [76] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6954–6963, 2021. 6, 7, 15, 16
- [77] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 88–107. Springer, 2020. 2, 4, 5, 6, 7, 9, 15, 16
- [78] Wenda Xu, Jia Pan, Junqing Wei, and John M Dolan. Motion planning under uncertainty for on-road autonomous driving. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2507–2512. IEEE, 2014. 2
- [79] Hiromitsu Yamada. Complete euclidean distance transformation by parallel operation. In *ICPR Proceedings*, pages 69–71, 1984. 2
- [80] Gengshan Yang and Deva Ramanan. Upgrading optical flow to 3D scene flow through optical expansion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1334–1343, 2020. 2
- [81] Jiaolong Yang, Hongdong Li, and Yunde Jia. Go-icp: Solving 3d registration efficiently and globally optimally. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1457–1464, 2013. 2

- [82] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. St3d: Self-training for unsupervised domain adaptation on 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10368–10378, 2021. 4
- [83] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. *arXiv preprint arXiv:2112.05131*, 2021. 2, 5
- [84] Alex Yu, Rui long Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 2
- [85] Jianqiao Zheng, Sameera Ramasinghe, Xueqian Li, and Simon Lucey. Trading positional complexity vs deepness in coordinate networks. In *European Conference on Computer Vision*, pages 144–160. Springer, 2022. 2, 5, 6, 13, 14

Appendix

A. Experiments

A.1. Speedup in network architectures

Positional encodings. Positional encodings (PEs) are usually used in coordinate networks [45] to increase the bandwidth of the input by multiple (random) frequencies. With PEs, the coordinate network converges much faster and achieves better performance [65]. Moreover, a recent paper [85] points out that the success of PEs is attributed to the rank increase of the input: the deepness of the neural network is to increase the rank of the embedding and aligns the embedding space to the output space. However, if the rank of the input is higher, the network can be shallower. Therefore, [85] uses a more complex PE to dramatically increase the rank of the input, thus the followed deep non-linear network can be replaced by a shallow linear function.

Complex PE-based linear model. In our paper, we implemented a complex positional encoding (PE) [85]-based linear model to test the efficiency of simplifying network architectures. While simple PE refers to a simple concatenation of the encoding in each input dimension, complex PE [85] is a more complicated encoding that computes the Kronecker product of the per-axis encoding. As mentioned in [85], one reason behind the success of the deep network is that it increases the rank of the low-rank input. Therefore, if the input to the network has a high rank, the network can be shallower accordingly. To increase the rank of the input, we use a complex encoding instead of a deep network. The rank of the complex encoding is

$$\begin{aligned} \text{Rank}(\phi(\mathbf{p}_x) \otimes \phi(\mathbf{p}_y) \otimes \phi(\mathbf{p}_z)) = \\ \text{Rank}(\phi(\mathbf{p}_x)) \text{Rank}(\phi(\mathbf{p}_y)) \text{Rank}(\phi(\mathbf{p}_z)), \end{aligned} \quad (8)$$

which achieves full rank that allows us to only use a linear layer \mathbf{W} as a follow-up embedder.

The advantage of a complex PE lies in two aspects: first, with a linear layer, the problem can be solved analytically in many cases; second, if the closed-form solution is difficult to obtain, using a linear layer in iterative solvers, such as gradient descent-based methods, will converge faster. Similar to frequency-based encodings [65], shift-based encodings like Gaussian or Triangle wave also work similarly well [85]. These shift-based encodings involve very few parameters for each sample point due to sparsity, while a deep network requires significant amounts of parameters.

To reconstruct the signal \mathbf{S} , we optimize

$$\arg \min_{\mathbf{W}} \left\| \text{vec}(\mathbf{S}) - (\phi(\mathbf{p}_x) \otimes \phi(\mathbf{p}_y) \otimes \phi(\mathbf{p}_z))^T \text{vec}(\mathbf{W}) \right\|_2^2. \quad (9)$$

When the coordinate is separable along each axis (*e.g.*, 2D image), using Kronecker product, we have a closed-form solution as

$$\mathbf{W} = \phi(\mathbf{p}_x)^{-1} \mathbf{S} \phi(\mathbf{p}_y)^{-T} \phi(\mathbf{p}_z)^{-T}. \quad (10)$$

With all the complex PE theory in hand, it is nontrivial to implement it in a scene flow problem, and to the best of our knowledge, we are the first to apply complex PE to real-world large-scale data. To employ complex PE in the scene flow problem, we first replace the non-linear multi-layer perceptrons (MLPs) in NSFP with a linear layer parameterized by $\mathbf{W} \in \mathbb{R}^{W_x W_y W_z \times 3}$. W_x, W_y , and W_z are encodings in each dimension and 3 is the dimension of the flow—and encode the input coordinates in complex PE form. The flow represented by MLPs can then be modified as

$$\mathbf{f} = g(\mathbf{p}; \mathbf{W}) = (\phi(\mathbf{p}_z) \otimes \phi(\mathbf{p}_y) \otimes \phi(\mathbf{p}_x)) \text{vec}(\mathbf{W}) \quad (11)$$

where $\phi(\cdot)$ is the encoder, $\mathbf{p}_x, \mathbf{p}_y, \mathbf{p}_z$ are the sample points in x, y, z coordinates.

Blending function. However, in the 3D point cloud case, the unordered points are not separable on each axis. A blending function B is introduced to interpolate the points and avoid the computation of large naive complex PE. Please note that the unordered point cloud has non-separable coordinates. According to [85], the non-separable-coordinate problem can be approximated by a blending function and an encoding of virtual separable grid points. The blending approximation is $\phi(\mathbf{p}_z) \otimes \phi(\mathbf{p}_y) \otimes \phi(\mathbf{p}_x) \approx B(\mathbf{p}; \phi) \phi(\mathbf{z}) \otimes \phi(\mathbf{y}) \otimes \phi(\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^{W_x}, \mathbf{y} \in \mathbb{R}^{W_y}$ and $\mathbf{z} \in \mathbb{R}^{W_z}$ are virtual grid points. Using such approximation, the computation of complex PE of grid points is as easy and fast as a matrix multiplication due to the property of the Kronecker product. Intuitively, the blending matrix B can be viewed as a matrix consisting of non-linear interpolation coefficients that depend on the encoding function $\phi(\cdot)$. It is large but sparse, *i.e.*, there are only 8 non-zero values on each row (corresponding to the 8 neighboring grid points of the query point), we can index the matrix efficiently by only querying the non-zero entries. Meanwhile, different from [85], grids here have physical meanings and their size can be adjusted.

Therefore, the scene flow becomes

$$\mathbf{f} \approx B(\mathbf{p}; \phi) \text{vec}(\phi(\mathbf{x}) \mathbf{W} \phi(\mathbf{z})^T \phi(\mathbf{y})^T), \quad (12)$$

where $\phi(\mathbf{x}) \mathbf{W} \phi(\mathbf{z})^T \phi(\mathbf{y})^T$ is a simple notation for n -mode multiplication. And the scene flow optimization is

$$\arg \min_{\mathbf{W}} \left\| \mathbf{f} - B(\phi(\mathbf{p}_x) \otimes \phi(\mathbf{p}_y) \otimes \phi(\mathbf{p}_z))^T \text{vec}(\mathbf{W}) \right\|_2^2. \quad (13)$$

We therefore solve \mathbf{W} using gradient descent and distance transform loss as

$$\begin{aligned} \mathbf{W}^* = \arg \min_{\mathbf{W}} & \sum_{\mathbf{p} \in \mathcal{S}_1} D(\mathbf{p} + B(\mathbf{p}; \phi) \\ & \text{vec}(\phi(\mathbf{x}) \mathbf{W} \phi(\mathbf{z})^T \phi(\mathbf{y})^T), \mathcal{S}_2). \end{aligned} \quad (14)$$

Compared to a L layer network of width W used in NSFP to process N sample points, a linear layer of size $W_x \times W_y \times W_z$ from the Kronecker product speeds up the network significantly from $\mathcal{O}(NW^2L)$ to $\mathcal{O}(8N + 3W_x W_y W_z (W_x + W_y + W_z))$. We further constrain the optimization by applying an explicit total variation (TV) regularizer on \mathbf{W} as:

$$\begin{aligned} \text{TV}(\mathbf{W}) = & \frac{1}{(W_x-1)(W_y-1)(W_z-1)} \\ & \sum_{i=0}^{W_x-2} \sum_{j=0}^{W_y-2} \sum_{k=0}^{W_z-2} \sqrt{(d\mathbf{x}_{i,j,k})^2 + (d\mathbf{y}_{i,j,k})^2 + (d\mathbf{z}_{i,j,k})^2}, \end{aligned} \quad (15)$$

where $d\mathbf{x}_{i,j,k} = \mathbf{W}_{i,j,k} - \mathbf{W}_{i+1,j,k}$, $d\mathbf{y}_{i,j,k} = \mathbf{W}_{i,j,k} - \mathbf{W}_{i,j+1,k}$, $d\mathbf{z}_{i,j,k} = \mathbf{W}_{i,j,k} - \mathbf{W}_{i,j,k+1}$. In all, the loss function of the complex PE model is (with TV)

$$\begin{aligned} \mathcal{L}(\mathbf{W}) = & \sum_{\mathbf{p} \in \mathcal{S}_1} D(\mathbf{p} + B(\mathbf{p}) \text{vec}(\phi(\mathbf{x}) \mathbf{W} \phi(\mathbf{z})^T \\ & \phi(\mathbf{y})^T), \mathcal{S}_2) + \frac{\lambda}{2} \text{TV}(\mathbf{W}). \end{aligned} \quad (16)$$

A.2. Speedup in point correspondence search

Before the deployment of the distance transform (DT), we explored other strategies to speed up the point correspondence search in Chamfer distance.

Build a k-d tree to search nearest neighboring points. The point distance function D is computationally intensive, as a set of point-to-point correspondences needs to be optimized in each optimization step. One speedup is to construct a k-d tree to accelerate the nearest neighbor search which reduces the computation complexity of point correspondence search from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \log n)$. Since the target point cloud \mathcal{S}_2 is fixed, we only need to pre-build the k-d tree for \mathcal{S}_2 once. However, the source point cloud \mathcal{S}_1 is deformed in each optimization step, making the pre-build of the source point cloud k-d tree happen in every iteration, not to mention that the per-iteration k-d tree query is another computation overhead when the number of points is big.

Randomly sample points. Stochastic gradient descent (SGD) [59] is now broadly used in large-scale learning problems. It approximates the actual gradient descent by

Table 4. Additional computation time and performance on Waymo Open Scene Flow dataset. The upper tabular between **blue bars** are experiments with the full point cloud, and the lower tabular between **orange bars** are experiments with only 8,192 points. Corr. / k-d tree / DT query denotes correspondence search, k-d tree-based correspondence search, or DT query.

Method	\mathcal{E}	Acc_5	Acc_{10}	θ_e	Pre-compute	$t (ms) \downarrow$		
	(m) \downarrow	(%) \uparrow	(%) \uparrow	(rad) \downarrow		Corr. / k-d tree / DT query	Network	Total
NSFP (baseline)	0.118	74.16	86.70	0.300	—	43.1 [15036]	2.38 [904]	18.39 s
Baseline (k-d tree CD)	0.104	74.13	86.81	0.296	15.24 [5283] 2.8x	2.27 [838] 1.05x	8.51 s 2.16x	
Baseline (k-d tree CD, linear)	0.101	70.14	86.24	0.315	12.92 [2349] 3.3x	1.39 [262] 1.71x	4.15 s 4.43x	
PointPWC-Net [77]	4.109	0.05	0.36	1.742	—	—	—	185 ms 1.32x
FlowStep3D [35]	0.753	0.01	0.09	1.212	—	—	—	725 ms 5.18x
PV-RAFT [76]	10.675	0.03	0.13	1.794	—	—	—	505 ms 3.61x
R3DSF [22]	0.414	35.47	44.96	0.527	—	—	—	140 ms
Ours (8,192 pts)	0.106	77.53	88.99	0.329	35.22	0.23 [6.5] 496x	2.60 [76] 1.8x	121 ms 1.16x

Table 5. Additional computation time and performance on Argoverse Scene Flow dataset.

Method	\mathcal{E}	Acc_5	Acc_{10}	θ_e	Pre-compute	$t (ms) \downarrow$		
	(m) \downarrow	(%) \uparrow	(%) \uparrow	(rad) \downarrow		Corr. / k-d tree / DT query	Network	Total
NSFP (baseline)	0.078	69.46	86.22	<u>0.253</u>	—	17 [5901]	2.31 [848]	8.38 s
Baseline (k-d tree CD)	0.078	69.14	85.99	0.253	11.3 [4063] 1.5x	2.28 [830] 1.0x	6.25 s 1.34x	
Baseline (k-d tree CD, linear)	0.071	68.72	86.39	0.288	9.36 [1701] 1.8x	1.41 [253] 1.6x	3.09 s 2.71x	
PointPWC-Net [77]	5.600	0.03	0.18	1.179	—	—	—	186 ms 1.65x
FlowStep3D [35]	0.845	0.01	0.08	1.860	—	—	—	729 ms 6.45x
PV-RAFT [76]	10.745	0.02	0.10	1.517	—	—	—	504 ms 4.46x
R3DSF [22]	0.417	32.52	42.52	0.551	—	—	—	113 ms
Ours (8,192 pts)	0.118	69.93	83.55	<u>0.352</u>	41.57	0.22 [6.33] 214x	2.51 [72.69] 1.9x	124 ms 1.10x

only computing the gradient of a randomly selected subset of the original dataset at each iteration. However, it can achieve relatively faster updates and guarantee a satisfied global convergence, especially when dealing with large-scale high-dimensional optimization [5].

Inspired by the idea of SGD, we choose to randomly subsample points of the dense point cloud at each iteration. Analogous to SGD, each individual point is viewed as sample data. A naive sampling strategy is to sample a fixed number of points. Instead, we develop sampling strategies based on the number of iterations or the decreasing percentage of the loss function. We sample fewer points at the beginning of the optimization when the point correspondences are noisy, and gradually increase the number of sampled points when the optimization becomes better constrained and finds better correspondences.

However, we also noticed that point sampling is not a practical strategy when applied to real-world problems, such as autonomous driving scenarios, where all points are needed to get sufficient information for detailed non-rigid motions.

Reduce the frequency of updating correspondence. Although Eq. (2) of the main paper optimizes network weights (scene flow) through an explicit point distance function, the point correspondence optimization is implicitly included. We have mentioned that the optimization of

the point correspondence and the scene flow are highly entangled. We cannot easily get a good scene flow estimation even given the optimal point correspondences.

Instead of separating the scene flow and correspondence optimization, we reduce the updating frequency of the point correspondences from every single iteration to several iterations. To guarantee a good initialization, we initially consecutively update correspondences for a fixed number of iterations.

However, the correspondence sampling strategy is unfavorable due to a considerable performance compromise.

A.3. Implementation details

We provide more implementation details for our method. Further details will be provided upon code release.

Datasets. We followed [38, 52] to create the pseudo scene flow labels, and removed ground points according to each dataset. Note that we used the raw point cloud from the lidar sensor and did not crop the data to a small range.

Truncated Chamfer distance. We used a truncated Chamfer distance loss for our baseline implementation as mentioned in the original NSFP [38] that is unbiased on extreme points. Practically, we chose $2m$ as a threshold to eliminate large point distance.

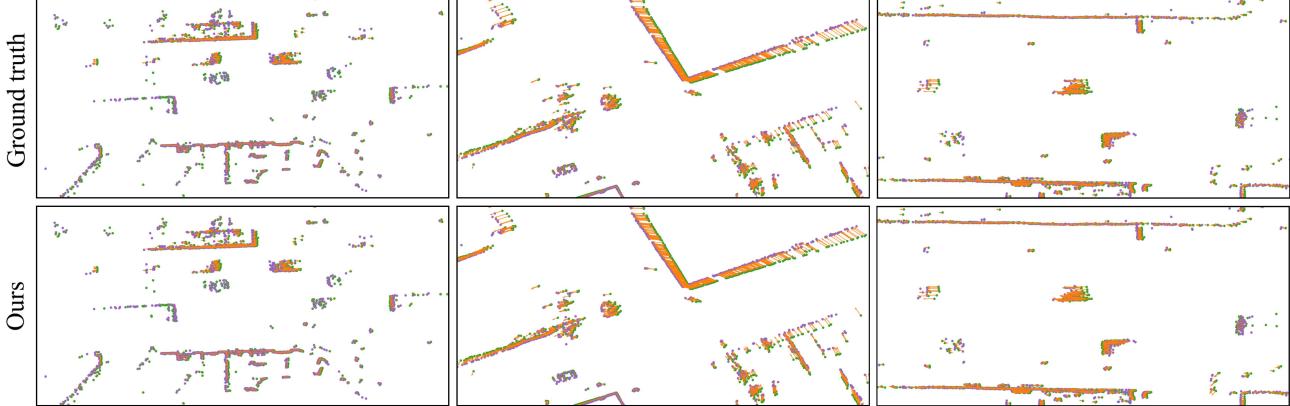


Figure 9. Visual results of the 2D scene flow estimation using our method on the Argoverse scene flow dataset. **Green** points are source, **purple** points are target, and **orange** arrows represent the flow vectors.

Complex positional encoding. Since point clouds are non-separable in 3D space, we first encoded the separable 3D virtual voxel vertices using shifted Gaussian encoders as depicted in Fig. 4 of the main paper. Since the 3D space in autonomous driving scenarios is large, we empirically found that a relatively larger voxel size (*e.g.*, 2m or 5m for autonomous driving scene flow datasets) that constrains the motions as rigid as possible within a larger local region is more suitable to encode scene flow. The choice of the Gaussian sigma also depends on the voxel size. Generally, the sigma of Gaussian encoding should be twice larger than the voxel size. For example, for a voxel size of 2m, $\sigma > 4$ is favored. Note that the Gaussian sigma and the voxel size can be adjusted within a small range.

A.4. Additional results

We provide additional results on Waymo Open and Argoverse scene flow datasets in Tab. 4 and Tab. 5 respectively.

We show how k-d tree-based correspondence search for CD loss speeds up the optimization, yet remains less effective, which indicates the inherent computation cost of correspondence search in CD loss cannot be easily solved using engineering techniques. The performance of the linear model drops by a large margin, suggesting that it is a less favorable choice for scene flow estimation.

A.5. Performance gap of learning methods

The performance of learning-based methods such as PointPWC-Net [77], FlowStep3D [35], PV-RAFT [76], FLOT [53] is inferior compared to non-learning-based methods (shown in Tab. 4 and Tab. 5, between the **orange** bar). As discussed in the main paper, the performance gap between the learning methods and the non-learning-based methods lies in the OOD generalizability. Even trained on similar lidar sensors—*e.g.*, FlowStep3D was

Table 6. Performance of distance transform with different grid cell sizes on a 2D BEV scene from Argoverse scene flow dataset. The total grid is of size 160×150.

Grid cell size (m)	$\mathcal{E} \downarrow$ (m)	$Acc_5 \uparrow$ (%)	$Acc_{10} \uparrow$ (%)	$\theta_\epsilon \downarrow$ (rad)	$t \downarrow$ (ms)	Mem. \downarrow (GB)
1	0.225	4.30	19.91	0.189	342	1.33
0.5	0.123	34.69	70.98	0.135	348	1.33
0.33	0.089	73.25	92.51	0.116	353	1.35
0.2	0.071	90.15	95.36	0.105	419	1.35
0.1	0.060	94.31	95.35	0.097	419	1.37
0.05	0.059	94.22	95.31	0.097	579	1.51
0.02	0.060	94.26	95.46	0.095	1151	2.68
0.01	0.058	93.92	95.25	0.095	1438	7.42

trained on the KITTI [44] dataset—the Waymo Open [64] and Argoverse [9] scene flow datasets have different point cloud range, coordinate configurations, *etc.* to KITTI dataset, making the pre-trained model vulnerable to these data variations. In contrast, non-learning-based methods maintained high accuracy on different datasets. Note that our method still has competitive efficiency among these learning methods.

A.6. Additional results of DT grid size

We provide additional results on a 2D bird’s eye view (BEV) scene in Tab. 6. The result is aligned with the main paper Fig. 7.

A.7. 2D BEV visual results

Some visual results of the 2D BEV scenes of the Argoverse scene flow dataset are shown in Fig. 9.

A.8. Visual results

Please see Fig. 10 and the project webpage <https://lilac-lee.github.io/FastNSF> for more visual results and applications.

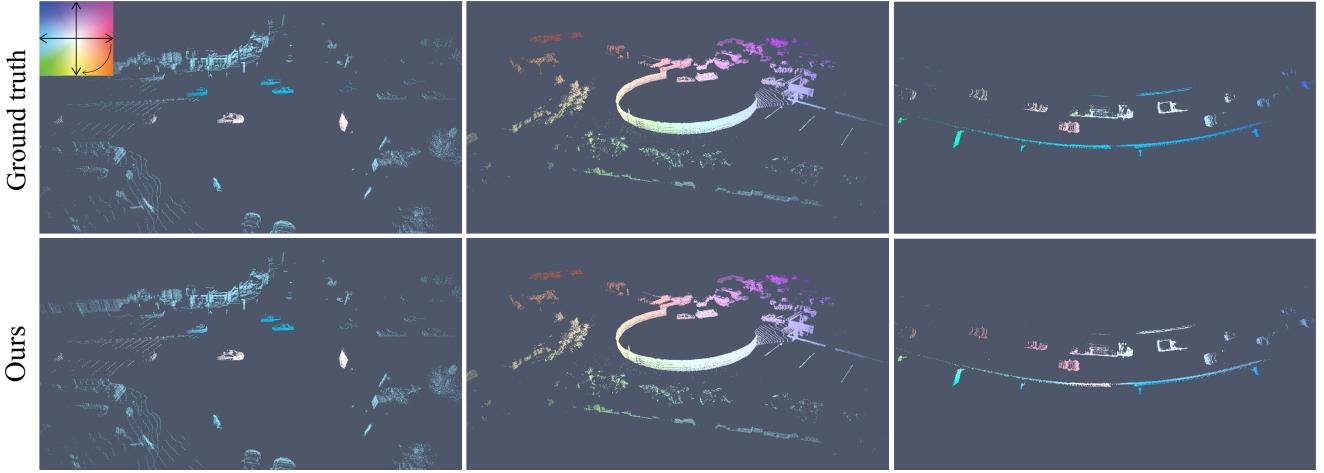


Figure 10. Visual examples of the scene flow prediction using our method on Waymo Open scene flow dataset.

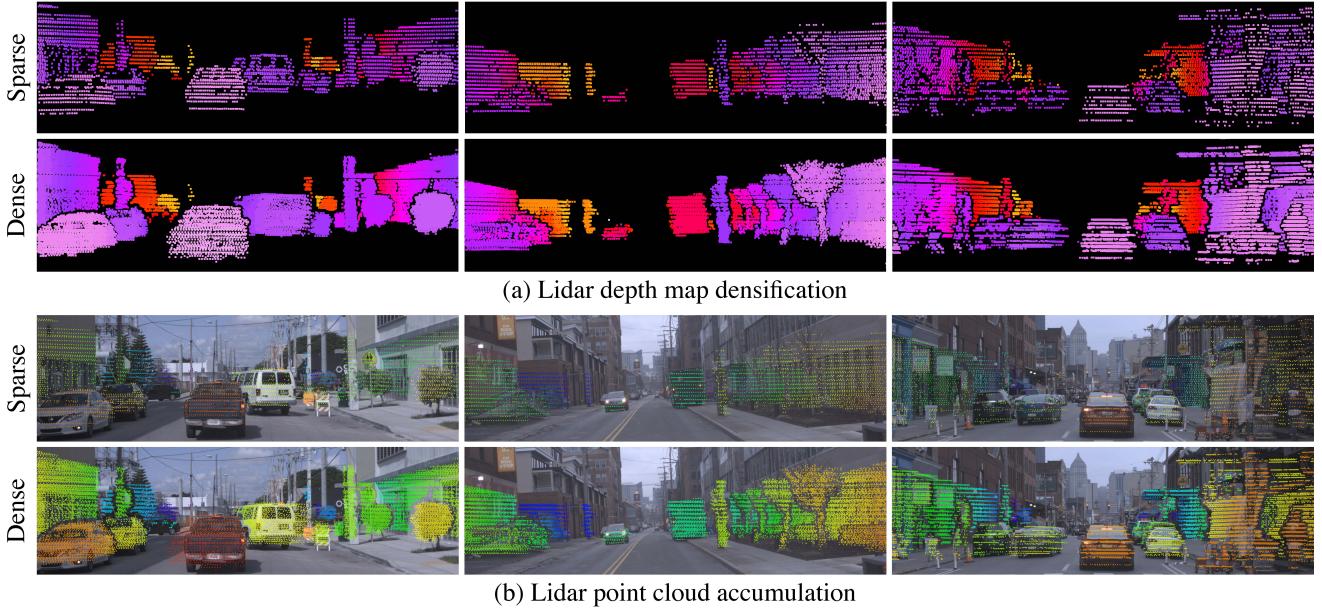


Figure 11. Visualization of the application of our method: (a) depth densification, and (b) point cloud accumulation. We present the original sparse scenes on the upper row and the densified results on the lower row. For point cloud accumulation, we projected the densified point cloud to the corresponding image plane for better visualization.

B. Application: point accumulation

The implicit and continuous neural function allows for easy point accumulation with per-pair scene flow estimation. Moreover, with the speedup that our method has achieved, the computation of point accumulation substantially decreased, making it possible for large amounts of point densification.

B.1. Continuous scene flow field

It is important to note that using DT to replace CD will not alter the continuous property. Therefore, similar to NSFP, our method creates a continuous flow field in that the network itself interpolates the motion of the entire 3D space, enabling

long-term flow estimation and point densification through forward integration.

B.2. Dense point cloud accumulation

We followed [38] to accumulate point clouds using Euler integration with per-pair scene flow estimation for the Argoverse scene flow dataset. Different from [38], we compute per-pair scene flow for each consecutive pair (*i.e.*, frame 1→2, frame 2→3, ..., frame 10→11) and interpolate fast neural scene flow to integrate 10 point clouds following the reference frame into the reference frame to densify the depth map and the point cloud. Some visual examples are shown in Fig. 11.