Fast Neural Scene Flow

Xueqian Li\(^{1}\)  Jianqiao Zheng\(^{1}\)  Francesco Ferroni\(^{2}\)  Jhony Kaesemodel Pontes\(^{3}\)  Simon Lucey\(^{1}\)

\(^{1}\)The University of Adelaide  \(^{2}\)NVIDIA  \(^{3}\)Latitude 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 [62] and Argoverse [8].

1. Introduction

Neural Scene Flow Prior (NSFP) [37] is considered the dominant method in open-world perception [47] and scene densification [37, 69] 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 [15, 31, 37, 47, 51].

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
which have gained significant speedups through architectural innovations [47]. 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) [17]. This differs considerably from other applications of coordinate networks throughout vision and learning such as neural radiance fields (NeRF [44]) and high-fidelity image reconstruction (NSFP) which have gained significant speedups through architectural innovations [83]. 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) [5, 13, 58] 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 [81]—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 [62] and Argoverse [8]. Our fast neural scene flow achieves up to $\sim 30$ times speedup than NSFP [37] (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 [67] uplifted from 2D optical flow. To reconstruct 3D flow, traditional image-based methods [2, 24, 25, 27, 28, 36, 53] formulate an optimization problem utilizing RGB or depth information, and learning-based RGB/RGB-D methods [6, 29, 30, 57, 60, 64, 78] 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, 11, 50] solve for a non-rigid registration problem, while recent work prefers full-supervised learning [23, 34, 38, 39, 52, 71, 73, 75] that uses point-based features or self-supervised learning [3, 21, 45, 65, 75] that employs a self-supervised loss. Recent non-learning-based methods [37, 51] draw our attention back to runtime optimization that easily scales to large data. Graph prior [51] explicitly builds a graph on the point cloud and uses a graph Laplacian regularizer. While neural scene flow prior [37] 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 [9, 20, 26, 46, 56, 81, 82] 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 [81] and TensorRF [9]. Recently, this trend was generalized for arbitrary signals through the introduction of complex positional encoding [83] 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, 5, 13, 41, 58] has played an important role in image processing, especially binary image analysis [48, 66]. Further applications are also found in medical image segmentation [12, 33, 59, 70, 72], robotics motion planning [55, 76], geometric representation [7, 10, 49], and accelerated point cloud registration [19, 79]. Among them, various distance measures have been used, such as city block, chessboard, and Euclidean distance [13, 77]. Naturally, Euclidean distance is preferred in computing point distance but it is also the most difficult metric to compute due to the temporal complexity [22]. Many work attempts to speedup Euclidean DT computation including raster-scan-based algorithms [5, 18, 35, 41, 54], fast marching-based algorithms [16, 40, 68], 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 $S_1$ (source) of the scene is sampled. Then given the movements of the sensor and objects in certain directions, another point cloud $S_2$ (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 $p \in S_1$ from time $t-1$ to time $t$ as a flow vector $f \in \mathbb{R}^3$, where $p' = p + f$. The translational vector set of all 3D points in $S_1$ is defined as the scene flow $\mathcal{F} = \{f_i\}_{i=1}^{|S_1|}$.

Therefore, the optimization of the scene flow is to minimize the point distance between the source $S_1$ and the target $S_2$. Usually, a regularization $C$, such as a Laplacian regularizer, is needed due to the highly unconstrained non-rigid flows. The overall optimization becomes

$$\mathcal{F}^* = \arg \min_{\mathcal{F}} \sum_{p \in S_1} D(p + f, S_2) + \lambda C, \quad (1)$$

where $D$ is a point distance function, $\lambda$ 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:

$$\Theta^* = \arg \min_{\Theta} \sum_{p \in S_1} D(p + g(p; \Theta), S_2). \quad (2)$$

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

Since the points in source $S_1$ and target $S_2$ are not in correspondence, nor having the same number of points, i.e., $|S_1| \neq |S_2|$, we use a distance function that handles these problems as

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

Practically, a bidirectional point distance is found, yielding the above equation equivalent to the Chamfer loss [17].

**Chamfer distance.** In point cloud processing, Chamfer distance (CD) [17] 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 can be written as

$$CD(S_1, S_2) = \sum_{p \in S_2} D(p, S_1) + \sum_{q \in S_1} D(q, S_2)$$

$$= \sum_{p \in S_2} \min_{x \in S_1} \|p - x\|^2 + \sum_{q \in S_1} \min_{y \in S_2} \|q - y\|^2. \quad (4)$$

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 $x \in S$ to the vertex $q$ of a DT map $\mathcal{G}$ as

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

We extend Eq. (5) to fit in the scene flow task such that $D$ refers to Euclidean distance, $x$ denotes the target point, and $q$ 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 single GPU. One strategy is to iteratively aggregate small...
Although we rasterize target points using a binary 3D image, an example of 2D DT map is shown in Fig. 2. Specifically, 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.

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 [12, 18], 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 [18] 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:

\[
L = \min_{y \in S', q \in G} \|y - q\|_2, \tag{6}
\]

where \(S'\) 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 [62] and Argoverse [8], 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 [37,51] and [31,80] respectively.

**Metrics** We follow scene flow metrics used in [37,38,45,51,75] to evaluate the performance. We also include the computation time breakdown in the table. These metrics are:

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

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

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

4) **angle error** \(\theta_e(rad)\) that measures the mean angle error (in radians) 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 [37]. 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 [34] and FLOT [52] are fully supervised methods trained on synthetic FlyingThings3D [42] datasets. PointPWC-Net [75] 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 [15, 31, 37, 47, 51]. 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 [21] is a weakly supervised method with no direct supervision of ground truth dynamic flows. We tested the method using the pretrained model (on KITTI [43]) 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.

4.1. Speedup in network architecture

Plenoxels [81] is a method that accelerates NeRF [44] 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) [83] to represent high-frequency signals and enable a linear reconstruction model which is similar in spirit to [81].

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 has been shown to improve the performance and convergence speed of coordinate networks [63]. 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

\[
\mathbf{f} \approx B(\mathbf{p}; \phi) \vec{\phi}(\mathbf{x}) \mathbf{W} \phi(\mathbf{z})^T \phi(\mathbf{y})^T,
\]

where \( B(\cdot) \) 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 interpolate 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 × numbers, and the slow-downs are marked with red ×. 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. ↑ means larger values are better while ↓ means smaller values are better. Corr. / DT query denotes correspondence search or DT query. Note: performance of learning methods PointPWC-Net [75], FlowStep3D [34], PV-RAFT [74] can be found in the supplementary material.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\mathcal{E}$ (m) ↓</th>
<th>Acc$_5$ (%) ↑</th>
<th>Acc$_10$ (%) ↑</th>
<th>$\theta$ (rad) ↓</th>
<th>t (ms) ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre-compute</td>
<td>Corr. / DT query</td>
<td>Network</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>NSFP [37]</td>
<td>0.100</td>
<td>76.62</td>
<td>88.56</td>
<td>0.286</td>
<td>114 [30972]</td>
</tr>
<tr>
<td>NSFP (baseline)</td>
<td>0.118</td>
<td>74.16</td>
<td>86.70</td>
<td>0.300</td>
<td>—</td>
</tr>
<tr>
<td>NSFP (baseline, linear)</td>
<td>0.096</td>
<td>70.78</td>
<td>86.31</td>
<td>0.310</td>
<td>9.48</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>0.072</strong></td>
<td><strong>84.73</strong></td>
<td><strong>92.24</strong></td>
<td><strong>0.280</strong></td>
<td>40.85</td>
</tr>
<tr>
<td>Ours (linear)</td>
<td>0.109</td>
<td>71.27</td>
<td>85.80</td>
<td>0.321</td>
<td>44.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLOT [52]</td>
<td>0.702</td>
<td>2.46</td>
<td>11.30</td>
<td>0.808</td>
<td>—</td>
</tr>
<tr>
<td>R3DSF [21]</td>
<td>0.414</td>
<td>35.47</td>
<td>44.96</td>
<td>0.527</td>
<td>—</td>
</tr>
<tr>
<td>NSFP [37] (8,192 pts)</td>
<td>0.138</td>
<td>53.62</td>
<td>78.57</td>
<td>0.339</td>
<td>—</td>
</tr>
<tr>
<td>Ours (8,192 pts)</td>
<td><strong>0.106</strong></td>
<td><strong>77.53</strong></td>
<td><strong>88.99</strong></td>
<td><strong>0.329</strong></td>
<td>35.22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>$\mathcal{E}$ (m) ↓</th>
<th>Acc$_5$ (%) ↑</th>
<th>Acc$_10$ (%) ↑</th>
<th>$\theta$ (rad) ↓</th>
<th>t (ms) ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre-compute</td>
<td>Corr. / DT query</td>
<td>Network</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>NSFP [37]</td>
<td><strong>0.069</strong></td>
<td>71.56</td>
<td>87.80</td>
<td><strong>0.235</strong></td>
<td>—</td>
</tr>
<tr>
<td>NSFP (baseline)</td>
<td>0.078</td>
<td>69.46</td>
<td>86.22</td>
<td>0.253</td>
<td>—</td>
</tr>
<tr>
<td>NSFP (baseline, linear)</td>
<td>0.097</td>
<td>67.03</td>
<td>83.20</td>
<td>0.314</td>
<td>9.19</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>0.071</strong></td>
<td><strong>80.05</strong></td>
<td><strong>90.71</strong></td>
<td><strong>0.289</strong></td>
<td>43.61</td>
</tr>
<tr>
<td>Ours (linear)</td>
<td>0.106</td>
<td>65.00</td>
<td>82.85</td>
<td>0.319</td>
<td>48.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLOT [52]</td>
<td>0.821</td>
<td>2.00</td>
<td>8.84</td>
<td>0.967</td>
<td>—</td>
</tr>
<tr>
<td>R3DSF [21]</td>
<td>0.417</td>
<td>32.52</td>
<td>42.52</td>
<td>0.551</td>
<td>—</td>
</tr>
<tr>
<td>NSFP [37] (8,192 pts)</td>
<td><strong>0.113</strong></td>
<td>46.32</td>
<td>72.68</td>
<td><strong>0.347</strong></td>
<td>—</td>
</tr>
<tr>
<td>Ours (8,192 pts)</td>
<td><strong>0.118</strong></td>
<td><strong>69.93</strong></td>
<td><strong>83.55</strong></td>
<td><strong>0.352</strong></td>
<td>41.57</td>
</tr>
</tbody>
</table>

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 [83], 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× speedup. However, the linear model only results in ~2× and ~1.2× 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) scene flow datasets.

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× 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× speedup was achieved while...
maintaining comparable high accuracy. However, replacing
da deep network with a linear model when using DT loss only
resulted in an additional 1.2/1.2 × 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 naïve point correspondence search, achieving a
remarkable 172/71 × speedup per optimization step—the
main contributor to the overall efficiency of our method.
Note that the pre-computing of the DT map (~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 × 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 [21, 34, 52, 74,
75]. For these learning models, a fixed number of points
is required and they only operate on fewer points, such as
2,048 [38, 52] or 8,192 points [21, 34, 52, 75]. 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 [32], 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 $160 \times 150$.

small subsets of dense point clouds using learning methods. However, this approach requires intensive computation, and it might cause out-of-memory issues [32].

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 [15, 31, 37, 47, 51]. 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 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 [15,31,37,47,51], 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>$E \downarrow$ (m)</th>
<th>$Acc_{5}$ ↑ (%)</th>
<th>$Acc_{10}$ ↑ (%)</th>
<th>$\theta_1 \downarrow$ (rad)</th>
<th>$t \downarrow$ (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>0.411</td>
<td>34.94</td>
<td>46.82</td>
<td>0.731</td>
<td>335</td>
</tr>
<tr>
<td>NSFP++ [47]</td>
<td>0.295</td>
<td>31.82</td>
<td>62.61</td>
<td>0.343</td>
<td>16188</td>
</tr>
<tr>
<td>NSFP++ (DT)</td>
<td>0.272</td>
<td>30.26</td>
<td>60.25</td>
<td>0.365</td>
<td>2001</td>
</tr>
</tbody>
</table>

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

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.
with our proposed distance transform (DT) loss, and we observed a robust and efficient surrogate to Chamfer distance loss. Therefore, we propose to use an efficient correspondence-free distance transform loss as a general loss that can be used in other 3D geometry tasks. Specifically, for optimization-based methods such as RSF [14], DT could replace the nearest neighbor distance loss to speed up the overall computation. For learning methods such as PointPWCNet [75], DT could replace Chamfer distance loss to speed up optimization during training, especially when dealing with a large number of points. For methods using cycle consistency, when replacing CD ($O(N^2)$) 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++ [47] 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 [8].

Following [47], 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.

4.4. 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 [14], DT could replace the nearest neighbor distance loss to speed up the overall computation. For learning methods such as PointPWCNet [75], DT could replace Chamfer distance loss to speed up optimization during training, especially when dealing with a large number of points. For methods using cycle consistency, when replacing CD ($O(N^2)$) 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++ [47] 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 [8].

Following [47], 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. 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 ($\sim 120$ 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++ [47].
References


[28] Frédéric Huguet and Frédéric Devernay. A variational method for scene flow estimation from stereo sequences. In


