

# Scene Flow from Point Clouds with or without Learning

Jhony Kaesemodel Pontes<sup>1</sup> James Hays<sup>1,2</sup> Simon Lucey<sup>3</sup>  
<sup>1</sup>Argo AI <sup>2</sup>Georgia Institute of Technology <sup>3</sup>The University of Adelaide

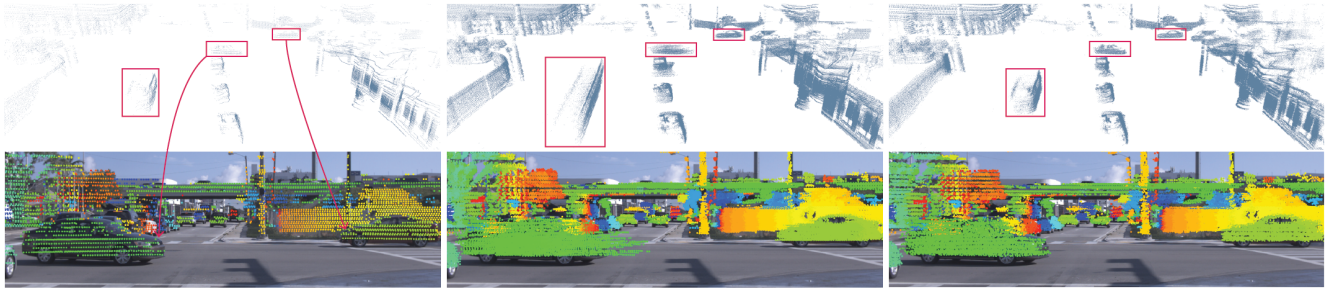


Figure 1. **Application of our scene flow approach for point cloud densification.** **Left:** Sparse point cloud and its projection onto an image. **Middle:** Densification using iterative closest point (ICP) which is a global rigid registration method. Five adjacent point clouds in each direction were used to densify the current frame. Note the “ghosting” effect due to the moving objects in the scene. **Right:** Densification using our method. Example from the ArgoVerse dataset. Colors repeat from green (close) to red (far) every fifty meters.

## Abstract

Scene flow is the three-dimensional (3D) motion field of a scene. It provides information about the spatial arrangement and rate of change of objects in dynamic environments. Current learning-based approaches seek to estimate the scene flow directly from point clouds and have achieved state-of-the-art performance. However, supervised learning methods are inherently domain specific and require a large amount of labeled data. Annotation of scene flow on real-world point clouds is expensive and challenging, and the lack of such datasets has recently sparked interest in self-supervised learning methods. How to accurately and robustly learn scene flow representations without labeled real-world data is still an open problem. Here we present a simple and interpretable objective function to recover the scene flow from point clouds. We use the graph Laplacian of a point cloud to regularize the scene flow to be “as-rigid-as-possible”. Our proposed objective function can be used with or without learning—as a self-supervisory signal to learn scene flow representations, or as a non-learning-based method in which the scene flow is optimized during runtime. Our approach outperforms related works in many datasets. We also show the immediate applications of our proposed method for two applications: motion segmentation and point cloud densification.

## 1. Introduction

Supervised learning approaches rely on a large amount of labeled data which is not always readily available. Also, the learned models are domain specific and do not generalize well to other scenarios that are statistically different from the training data. For example, a model trained to predict scene flow on an unrealistic synthetic dataset will likely not perform well on a real self-driving situation. Current supervised methods that estimate scene flow from point clouds [13, 18, 41] are trained on large-scale synthetic data, e.g. FlyingThings3D [19], to be further fine-tuned on small real-world datasets such as the KITTI Scene Flow [20, 22]. Although there is an increasing availability of large-scale self-driving datasets with point cloud data from light detection and ranging sensors (lidar) [7, 9, 31], they do not provide scene flow labels. Annotation of scene flow on lidar data is challenging and expensive, given that a pair of point clouds needs translational vector labels describing the motion at every point in the scene.

As a result, self-supervised methods have recently been proposed [23, 45] to learn scene flow representations without leveraging human supervision. Given two consecutive point clouds, the idea is to estimate how each point in the source point cloud moves towards its corresponding point on the next point cloud. The collection of all the individual motions, or translational vectors, is the scene flow. A global rigid transformation to represent dynamic scene flow is not sufficient—instead there is a transformation for every point. However, solving for individual transformations

is an ill-posed problem [1]. Most of the prior work aims at improving scene flow estimation by using deep hierarchical networks [13, 18, 23, 41, 45]. They perform recursive sampling and grouping of neighboring points in different scales to regularize the scene flow to be similar in local regions.

In this work, we propose a simple and geometrically interpretable objective function to optimize the scene flow. We constrain the scene flow such that points that are close to each other in a local region move rigidly while the collection of all points move “as-rigid-as-possible” [29]. Such intuition is elegantly captured by the widely used graph Laplacian [5, 11, 17, 27–29, 39] which implicitly embeds the topological structure of a point cloud. Therefore, we propose the formation of an explicit graph on the source point cloud to capture its topology and context information about how points are locally connected. The graph Laplacian representation allows us to regularize and robustly approximate the scene flow from point clouds without annotations.

Our proposed objective function can be used for self-supervised learning or for non-parametric optimization (for clarity, we exchange *non-parametric* to *non-learning* from now on). We evaluate our method on the synthetic FlyingThings3D [19] dataset and on the real-world KITTI Scene Flow [20, 22], Argoverse [9], and nuScenes [7] datasets. Since Argoverse and nuScenes do not provide scene flow annotations, we use the ego-vehicle poses and 3D object tracks to create pseudo labels.

Our approach outperforms prior self-supervised works on many datasets *with or without learning*. It not only provides a robust self-supervisory signal to learn scene flow representation but is also a robust non-learning objective for runtime optimization. Furthermore, we explore the applications of our method for two applications: motion segmentation and point cloud densification.

**Our main contributions are:**

- a simple and geometrically interpretable objective function to approximate scene flow from a pair of point clouds. An “as-rigid-as-possible” regularizer constrains the non-rigid motion flow based on the graph Laplacian of the source point cloud;
- our objective function can be used with or without learning. As a self-supervisory signal to learn scene flow representations, or as a non-learning method in which the scene flow is optimized during runtime;
- we show compelling results on synthetic, FlyingThings3D, and real-world datasets, KITTI Scene Flow, Argoverse, and nuScenes—with or without learning.

## 2. Related Work

Here we review the most relevant scene flow works based on two broad categories: non-learning-based and learning-based scene flow methods. We focus on point-based scene

flow, but for completeness we also include a review of image-based scene flow methods. Moreover, we briefly review related works that make use of the graph Laplacian.

### 2.1. Non-learning-based scene flow

**Image-based scene flow.** The development of non-learning scene flow methods is traced back to the classical work of Vedula *et al.* [35]. They represented the scene flow as a dense vector field defined for every point on every surface in the scene. Then, a two-step approach was proposed to estimate the scene flow in a decoupled way: (1) optical flow is computed for the multi-view image sequence; (2) since the optical flow is the projection of the scene flow onto the image plane, the scene flow is reconstructed through triangulation using the optical flow information. Following Vedula *et al.*’s work and building upon developments in optical flow and stereo matching, other works proposed the use of variational methods to jointly estimate the scene flow and the 3D structure from stereo sequences [2, 14, 25, 34, 42, 43]. Other relevant works relied on rigidity assumptions of the scene structure and motion [21, 36–38].

**Point-based scene flow.** Most non-learning scene flow methods are image-based. Nevertheless, we can interpret the seminal non-rigid iterative closest point (NICP) work by Amberg *et al.* [1] and related non-rigid registration methods as scene flow estimators. NICP is an iterative method to deform a 3D template to fit scanned meshes. The algorithm uses a locally affine regularization based on the mesh topology to constrain the deformation field to be smooth. However, NICP relies on proper initialization, is sensitive to holes in the geometry, and is only suitable for small-scale differences between the template and the scanned mesh.

### 2.2. Learning-based scene flow

**Image-based scene flow.** Learning-based methods to estimate scene flow from monocular images have recently been proposed [6, 15, 46]. Since the monocular scene flow problem is ill-posed by nature, most methods rely on 3D prior assumptions learned from data. Yang and Ramanan [46] proposed a learning method to lifting optical flow to scene flow from monocular images using optical expansion.

**Point-based scene flow.** There has been a great interest in estimating scene flow directly from point clouds. Recently, Dewan *et al.* [10] proposed to estimate rigid scene flow from point clouds. They proposed an energy minimization problem of a factor graph with hand-crafted signature of histograms of orientations (SHOT) descriptors [32] for correspondence search. Later, Ushani *et al.* [33] proposed a logistic classifier to predict if two columns of occupancy grids are in correspondence. Then, they formulated an expectation-maximization (EM) algorithm to estimate a locally rigid scene flow. Behl *et al.* [3] proposed a method to jointly predict scene flow and 3D bounding boxes with their

rigid body motions. Liu *et al.* [18] proposed FlowNet3D to extract point features using PointNet++ [26] and to estimate the scene flow using a flow embedding layer. Wang *et al.* [41] proposed FlowNet3D++ to improve on FlowNet3D by incorporating geometric constraints in the form of point-to-plane distance and angular alignment. Gu *et al.* [13] presented HPLFlowNet to predict the scene flow using bilateral convolutional layers (BCL) to project the point cloud onto a permutohedral lattice. Wu *et al.* [45] proposed PointPWCNet, a coarse-to-fine approach to predict scene flow from point clouds. Mainly, they employed PointConv [44] to learn features from points efficiently. Mittal *et al.* [23] presented a method to fine-tune a pre-trained FlowNet3D model with real data using a self-supervised loss based on nearest neighbors and cycle consistency.

### 2.3. Graph Laplacian-based methods

The graph Laplacian is a basis for a variety of geometry processing tasks, and it has been widely used for mesh editing [5, 17, 28, 29], mesh registration [11], physics-based modeling [27], spectral graph analysis [39], among others. It is generally applied to achieve a common goal: constrain the parameters of deformation to be smooth functions along a surface. However, the graph Laplacian has not yet been fully investigated for the point-based scene flow problem.

### 2.4. Overview

Current point-based scene flow approaches rely on deep networks to achieve state-of-the-art performance over classical methods. They are usually supervised from large-scale synthetic data and then fine-tuned on small-scale real-world datasets. Given the scarcity of large-scale, real-world data with scene flow annotations, recent works have shifted towards self-supervised methods. To regularize the scene flow, these methods perform iterative sampling and grouping of neighboring point features in different scales. Our approach differs from the previously mentioned works in that our objective function does not rely on recursive point feature sampling and grouping as in [23, 45]. Instead, we build an *explicit* graph on the source point cloud to capture the topology and context information about how points are locally connected and how they should move through the graph Laplacian. We seek to explore traditional objectives that have been abandoned for supervised methods.

## 3. Background

In this section, we briefly summarize the graph Laplacian formation process. Let  $\mathbf{G}=\{\mathbf{V}, \mathbf{E}\}$  be an undirected graph with a set of vertices  $\mathbf{V}$  connected by a set of edges  $\mathbf{E}$ . For a given set of  $\mathbf{V}$  and  $\mathbf{E}$ , the graph can be formally represented by its adjacency matrix  $\mathbf{A} \in \mathbb{R}^{N \times N}$  which describes the vertex connectivity for  $N$  vertices. The element  $A_{ij}$  of  $\mathbf{A}$  assume values  $\{0, 1\}$ . The value  $A_{ij}=0$  is assigned if the

vertices  $i$  and  $j$  are not connected with an edge, and  $A_{ij}=1$ , if these vertices are connected, that is

$$A_{ij} \stackrel{\text{def}}{=} \begin{cases} 1, & \text{if } (i, j) \in \mathbf{E} \\ 0, & \text{if } (i, j) \notin \mathbf{E}. \end{cases} \quad (1)$$

The weight matrix  $\mathbf{W}$  is similar to the adjacency matrix  $\mathbf{A}$  definition but it can convey extra contextual information, for example, the relative importance of the edge connections based on a distance metric. Another matrix that captures information from a graph is the degree matrix  $\mathbf{D}$ . For an undirected graph,  $\mathbf{D}$  is a diagonal matrix whose elements  $D_{ii}$  are equal to the sum of weights of all edges connected to the vertex  $i$ , that is, the sum of elements in its  $i$ -th row

$$D_{ii} \stackrel{\text{def}}{=} \sum_{j=1}^{N-1} W_{ij}. \quad (2)$$

The graph Laplacian matrix  $\mathbf{L}$  combines the weight matrix and the degree matrix as  $\mathbf{L} \stackrel{\text{def}}{=} \mathbf{D} - \mathbf{W}$ .

The elements of  $\mathbf{L}$  are nonnegative real numbers at the diagonal positions and nonpositive real numbers at the off-diagonal positions. For an undirected graph,  $\mathbf{L}$  is symmetric, *i.e.*  $\mathbf{L}=\mathbf{L}^T$ . For practical reasons, it is often advantageous to use a normalized  $\mathbf{L}$  [30], defined as

$$\mathbf{L} \stackrel{\text{def}}{=} \mathbf{D}^{-1/2}(\mathbf{D} - \mathbf{W})\mathbf{D}^{-1/2}. \quad (3)$$

The graph Laplacian  $\mathbf{L}$  indicates how smooth a graph function is. It is the discrete version of the Laplacian for continuous spaces which is defined by the second derivative of a function. A smooth graph function does not drastically change in value from one vertex to another connected vertex. Thus, we seek to explore  $\mathbf{L}$  as a regularization term to enforce an “as-rigid-as-possible” [29] scene flow.

## 4. Method

We introduce a method to estimate the scene flow given two consecutive point clouds from a dynamic scene. We constrain the scene flow with the graph Laplacian such that points that are close to each other move rigidly while the collection of all points move “as-rigidly-as-possible”.

### 4.1. Problem formulation

Let  $\mathbf{P}_{t-1} \in \mathbb{R}^{n_1 \times 3}$  be the 3D point cloud with  $n_1$  points at time  $t-1$  (source point cloud) and  $\mathbf{P}_t \in \mathbb{R}^{n_2 \times 3}$  be the point cloud with  $n_2$  points at time  $t$  (target point cloud). To recover the scene flow, the source point cloud  $\mathbf{P}_{t-1}$  should move close to the target point cloud  $\mathbf{P}_t$ . Therefore, a criterion for the fitting is that each point  $\mathbf{p}_{t-1} \in \mathbf{P}_{t-1}$  should be as near as possible to its corresponding point in  $\mathbf{P}_t$ . In a rigid scene, a global rigid transformation is sufficient to recover the scene flow—all points  $\mathbf{p}_{t-1}$  share the same transformation. However, most scenes we are interested in are non-rigid. For example, objects within a self-driving scenario are dynamic and have independent behaviors. A

formulation to take into account all non-rigid motions is to solve for a translational vector  $\mathbf{f} \in \mathbb{R}^3$  for each point  $\mathbf{p}_{t-1} \in \mathbf{P}_{t-1}$ . The collection of all translational vectors  $\mathbf{f}$  is the scene flow  $\mathbf{F} \in \mathbb{R}^{n_1 \times 3}$ . Thus  $\mathbf{P}_{t-1} + \mathbf{F}$  translates  $\mathbf{P}_{t-1}$  towards  $\mathbf{P}_t$ . If the scene flow is projected onto an image plane the optical flow is granted for free.

## 4.2. Data term

An optimal scene flow estimation means an exact match between the source point cloud  $\mathbf{P}_{t-1}$  and the target point cloud  $\mathbf{P}_t$ . However, real-world point clouds do not necessarily have the same number of points ( $n_1$  is likely different from  $n_2$ ) nor exact correspondences for exact matching. Thus, we define a data objective term  $E_d$  to measure the quality of the point cloud matching as

$$E_d(\mathbf{F}) = \sum_{i=1}^{n_1} \text{dist}^2(\mathbf{p}_{t-1_i} + \mathbf{f}_i, \mathbf{P}_t), \quad (4)$$

where the  $\text{dist}(\cdot)$  function computes the distance to the closest corresponding point on  $\mathbf{P}_t$ . If all of the  $\mathbf{f}_i$  are constrained to be the same, then minimizing the data term would solve for a global rigid transformation.

## 4.3. Graph Laplacian term

Each displacement vector  $\mathbf{f}_i$  has three degrees of freedom in the optimization. If only using the data term  $E_d$ , the problem is underconstrained as there are as many translations as there are points in the source point cloud. To constrain the problem, we solve for a set of transformations that are ‘‘as-rigid-as-possible’’. We formulate the constraint as

$$E_{\mathcal{L}}(\mathbf{F}) = \sum_{\{i,j\} \in \mathbf{E}} \|\mathbf{f}_i - \mathbf{f}_j\|_2^2, \quad (5)$$

where  $\mathbf{E}$  is the set of edges of a graph  $\mathbf{G}$  formed on  $\mathbf{P}_{t-1}$ .

There exist a variety of choices in which graphs are constructed from a set of points, including  $k$ -nearest neighbor ( $k$ -NN) graphs,  $r$ -neighborhood graphs, and ‘‘self-tuning’’ graphs [47]. However, the two most commonly used are the  $r$ -neighborhood graph and the  $k$ -NN graph. In the  $r$ -neighborhood graph, the neighborhoods are restricted to lie inside a sphere with radius  $r$ . The  $k$ -NN graph does not define the neighborhoods with a fixed length-scale  $r$  but rather by specifying for each point a set of  $k$  nearest neighbors.

In this work, we focus on  $k$ -NN graphs as they are almost always preferred in practice over  $r$ -neighborhood graphs due to its simplicity and better sparsity and connectivity properties [8]. Hence we construct a  $k$ -NN graph  $\mathbf{G}$  on the source point cloud  $\mathbf{P}_{t-1}$  with the purpose of leveraging its geometrical and topological information. Formally, given the set of  $n_1$  points  $\mathbf{P}_{t-1} = \{\mathbf{p}_{t-1_1}, \dots, \mathbf{p}_{t-1_{n_1}}\}$ , the undirected  $k$ -NN graph consists of the vertex set  $\mathbf{V}$  and the edge set  $\mathbf{E}$  which is a subset of  $\mathbf{P}_{t-1} \times \mathbf{P}_{t-1}$ . The vertices  $\mathbf{p}_{t-1_i}$  and  $\mathbf{p}_{t-1_j}$  are linked to form an edge if  $\mathbf{p}_{t-1_i}$  is a  $k$ -nearest neighbor of  $\mathbf{p}_{t-1_j}$  or vice versa.

Given  $\mathbf{G}$ , we compute the normalized graph Laplacian  $\mathbf{L} \stackrel{\text{def}}{=} \mathbf{D}^{-1/2}(\mathbf{D} - \mathbf{W})\mathbf{D}^{-1/2}$ , where  $\mathbf{D}$  is the degree matrix and  $\mathbf{W}$  is the weight matrix with elements  $W_{ij}$  defined as

$$W_{ij} = \begin{cases} e^{-r_{ij}^2}, & \text{if } (i, j) \in \mathbf{E}, \\ 1, & \text{if } i = j, \\ 0, & \text{if } (i, j) \notin \mathbf{E}, \end{cases} \quad (6)$$

where  $r$  is the Euclidean distance between the vertices  $i$  and  $j$  of  $\mathbf{G}$ . We use a weighted graph such that the edges convey information about the relative importance of their connection based on its distance. Finally, the term on Eq. 5 can be redefined in matrix form using the graph Laplacian  $\mathbf{L}$  as

$$E_{\mathcal{L}}(\mathbf{F}) = \text{tr}(\mathbf{F}^T \mathbf{L} \mathbf{F}), \quad (7)$$

where the symbol  $\text{tr}(\cdot)$  is the trace of a matrix.

The full objective function is defined as

$$E(\mathbf{F}) = E_d + \alpha E_{\mathcal{L}}, \quad (8)$$

where the first term minimizes the distance between the transformed source point cloud  $\mathbf{P}_{t-1} + \mathbf{F}$  and the target  $\mathbf{P}_t$ . The idea is to search for the scene flow  $\mathbf{F}$  that best transforms  $\mathbf{P}_{t-1}$  towards  $\mathbf{P}_t$ . The second term is the graph Laplacian constraint meaning that the objective function should not change too much between nearby points, *i.e.* ‘‘as-rigid-as-possible’’, and  $\alpha$  is a weighting factor for the graph Laplacian regularizer. Note that the graph Laplacian  $\mathbf{L}$  is constant throughout the minimization.

## 4.4. Point cloud correspondences

The data term  $E_d$  relies on a function  $\text{dist}(\cdot)$  to compute the distance to the closest corresponding point on  $\mathbf{P}_t$ , *i.e.* for every point on  $\mathbf{P}_{t-1}$ , the function has to find the closest point on  $\mathbf{P}_t$ . An inexpensive way to approximate the correspondences between point clouds is through the Chamfer distance function [12]. It is the average matching distance to the nearest points, and it is defined as

$$\mathcal{C}(\mathbf{P}_t, \mathbf{P}_{t-1}) \stackrel{\text{def}}{=} \sum_{\mathbf{x} \in \mathbf{P}_t} \min_{\mathbf{y} \in \mathbf{P}_{t-1}} \|\mathbf{x} - \mathbf{y}\|_2 + \sum_{\mathbf{y} \in \mathbf{P}_{t-1}} \min_{\mathbf{x} \in \mathbf{P}_t} \|\mathbf{x} - \mathbf{y}\|_2. \quad (9)$$

For each point, the Chamfer distance finds the nearest neighbor in the other point cloud and sums the squared distances. Since it is a function of point locations in  $\mathbf{P}_t$  and  $\mathbf{P}_{t-1}$ , the Chamfer distance is a continuous and a piecewise smooth function. Hence differentiable almost everywhere and applicable as a loss function [12].

The proposed objective function can be optimized through gradient-descent algorithms. Thus it can be used *with or without learning*. With learning, one can employ the proposed objective as a loss function for self-supervised learning. Without learning, one can recover scene flow directly from pairwise point clouds without any supervision.

## 5. Experiments

### 5.1. Setup

**Datasets.** We used the following four datasets:

**1. FlyingThings3D** [19] is a large-scale synthetic dataset consisting of stereo and RGB-D images rendered from randomly moving 3D CAD objects. We used the preprocessed dataset released by [18], where the RGB-D images were converted to point clouds and the optical flow to scene flow. It contains 19,967 train samples and 2,000 test samples.

**2. KITTI Scene Flow** [20, 22] was designed to evaluate image-based scene flow methods on self-driving scenarios. Lidar point clouds were collected using the Velodyne HDL-64E sensor. They accumulated seven nearby point clouds and projected onto the images to densify the depth maps. We used the preprocessed dataset released by [18], where they lifted the depth maps to points clouds and the optical flow to scene flow. It contains 100 train and 50 test samples.

**3. Argoverse** [9] is a new self-driving dataset, in the spirit of KITTI, but with more data and HD maps containing lane centerlines and ground height. However, scene flow annotations are not provided. To quantitatively evaluate our method, we created a dataset, “*Argoverse Scene Flow*”, based on the information provided in the Argoverse 3D Tracking v1.1 set. Specifically, we used the point clouds sensed from two Velodyne VLP-32 sensors, the vehicle poses and the 3D object tracks to lift pseudo scene flow annotations<sup>1</sup>. It contains 2,691 train and 212 test samples.

**4. nuScenes** [7] is a large-scale self-driving dataset featuring tracking annotations, map information, lidar point clouds collected with a Velodyne VLP-32 sensor, among others. However, as in the Argoverse dataset, scene flow annotations are not provided. We use the same Argoverse preprocessing steps to create the “*nuScenes Scene Flow*”<sup>1</sup>. It contains 1,513 train samples and 310 test samples.

For a fair comparison with previous works, we also removed the ground points from the Argoverse Scene Flow and nuScenes Scene Flow datasets<sup>1</sup>. Moreover, we noted that Argoverse and nuScenes might contain many rigid scenes where ICP-based methods would perform well on them. However, to avoid biasing to rigid scene flows, we filtered rigid scenes from Argoverse and nuScenes.

**Metrics.** We used the following metrics: **1.**  $\mathcal{E}$ : the mean absolute distance error in meters, or end-point error; **2.**  $\%_5$ : the percentage of flow vectors where  $\mathcal{E} < 0.05$  or  $\delta < 5\%$ ;  $\delta$  is the percent error; **3.**  $\%_{10}$ : the percentage of flow vectors where  $\mathcal{E} < 0.1$  or  $\delta < 10\%$ ; and **4.**  $\theta_e$ : the mean angle error in radians between the estimated and ground-truth scene flow.

**Implementation details.** We used the automatic differentiation in PyTorch [24] to optimize our objective function using Adam [16]. In the “*with learning*” setting, we trained

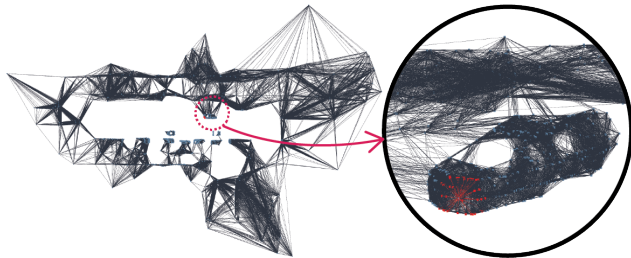


Figure 2. **Implementation details.** Example of a  $k$ -NN graph formed on a point cloud from Argoverse Scene Flow. **Left:** Top view of the point cloud with its graph. **Right:** Zoomed-in area shows a vehicle and a local neighborhood region illustrated in red. The graph was built using fifty nearest neighbors (*i.e.*  $k=50$ ).

FlowNet3D on FlyingThings3D with supervision for 300 epochs, batch size of 16, and a learning rate  $\gamma$  of  $1e-3$  and decaying it every 100 epochs by a factor of 0.1. We further fine-tuned the model with the self-supervision of our proposed objective function on each real-world dataset. To avoid insufficient training, we used 1,000 epochs for fine-tuning,  $\gamma$  of  $1e-5$ , and the graph Laplacian regularizer weight  $\alpha$  of 4.0 (found by grid-search). In the “*without learning*” setting, we ran the optimization for 1,500 epochs starting with a  $\gamma$  of 0.1 and an  $\alpha$  of 10.0. We set the number of neighbors  $k$  to 50 to construct the  $k$ -NN graph. We performed experiments on point clouds with 2,048, 4,096, and 8,192 points. Fig. 2 shows an example of a graph formed on a point cloud from Argoverse Scene Flow.

**Runtime.** It took  $\sim 15$  h to train FlowNet3D on FlyingThings3D for 300 epochs and  $\sim 3$  h to fine-tune it on KITTI Scene Flow with our self-supervised loss for 1,000 epochs. The fine-tuning on Argoverse Scene Flow took  $\sim 3$  days and on nuScenes took  $\sim 2$  days, both for 1,000 epochs. At test time, it took  $\sim 36$  ms for a single prediction using the learning-based approach. The non-learning version took  $\sim 9$  s to optimize a single scene flow. Experiments were run with 2,048 points, and on an NVIDIA Quadro P5000 GPU.

### 5.2. Results

Table 1 compared our method to learning-based methods (*with learning*): with supervision and with self-supervision, and against non-learning methods (*without learning*).

#### 5.2.1 With learning

**With supervision.** We tested the generalization capabilities of three off-the-shelf state-of-the-art supervised methods: Deep Closest Point (DCP) [40], FlowNet3D [18], and PointPWC-Net [45]. DCP predicts a *rigid* scene flow (*i.e.* a global transformation). However, we used it as a baseline to verify that rigid approaches are not outperforming the non-rigid methods. We used the pre-trained DCP model on the ModelNet40 dataset and tested it on different datasets. Since DCP was trained on noise-free CAD objects, it could

<sup>1</sup>Please refer to the supplementary material for details.

	FlyingThings3D [19]				KITTI Scene Flow [18]				Argoverse Scene Flow [9]				nuScenes Scene Flow [7]			
	#Train: 19,967   #Test: 2,000				#Train: 100   #Test: 50				#Train: 2,691   #Test: 212				#Train: 1,513   #Test: 310			
	$\mathcal{E} \downarrow$ (m)	$\%_5 \uparrow$ (%)	$\%_{10} \uparrow$ (%)	$\theta_e \downarrow$ (rad)	$\mathcal{E} \downarrow$ (m)	$\%_5 \uparrow$ (%)	$\%_{10} \uparrow$ (%)	$\theta_e \downarrow$ (rad)	$\mathcal{E} \downarrow$ (m)	$\%_5 \uparrow$ (%)	$\%_{10} \uparrow$ (%)	$\theta_e \downarrow$ (rad)	$\mathcal{E} \downarrow$ (m)	$\%_5 \uparrow$ (%)	$\%_{10} \uparrow$ (%)	$\theta_e \downarrow$ (rad)
• Deep Closest Point (DCP) [40]	1.007	0.17	0.96	1.213	0.638	1.84	3.91	0.708	1.236	0.09	0.51	1.453	1.277	0.10	1.04	1.434
• FlowNet3D [18]	0.134	22.64	54.17	0.305	0.199	10.44	38.89	0.386	0.455	1.34	6.12	0.736	0.505	2.12	10.81	0.620
• PointPWC-Net [45]	<b>0.121</b>	<b>29.09</b>	<b>61.70</b>	<b>0.229</b>	<b>0.142</b>	<b>29.91</b>	<b>59.83</b>	<b>0.239</b>	<b>0.405</b>	<b>8.25</b>	<b>25.47</b>	<b>0.674</b>	<b>0.442</b>	<b>7.64</b>	<b>22.32</b>	<b>0.497</b>
• Just Go with the Flow [23]			—		0.218	10.17	34.38	0.254	0.542	8.80	20.28	0.715	0.625	6.09	0.139	0.432
• PointPWC-Net (self-sup. loss) [45]			—		0.177	13.29	42.15	0.272	0.409	9.79	<b>29.31</b>	0.643	0.431	6.87	22.42	0.406
• Ours			—		<b>0.169</b>	<b>21.71</b>	<b>47.75</b>	<b>0.254</b>	<b>0.353</b>	<b>12.90</b>	28.33	<b>0.604</b>	<b>0.284</b>	<b>14.50</b>	<b>35.46</b>	<b>0.363</b>
• Iterative Closest Point (ICP) [4]	0.412	16.87	34.56	0.605	0.409	5.24	28.14	0.608	0.438	8.50	24.70	0.665	0.380	15.03	34.78	0.450
• Non-rigid ICP (NICP) [1]	0.339	14.05	35.68	0.480	0.338	22.06	43.03	0.460	0.461	4.27	13.90	0.741	0.402	6.99	21.01	0.492
• PointPWC-Net (self-sup. loss) [45]	0.433	6.23	19.46	0.643	0.272	16.98	35.65	0.338	0.466	8.41	22.62	0.701	0.399	8.31	23.30	0.454
• Ours	<b>0.259</b>	<b>16.30</b>	<b>41.60</b>	<b>0.369</b>	<b>0.093</b>	<b>64.76</b>	<b>82.13</b>	<b>0.137</b>	<b>0.257</b>	<b>25.26</b>	<b>47.50</b>	<b>0.467</b>	<b>0.288</b>	<b>20.19</b>	<b>43.59</b>	<b>0.337</b>

Table 1. **Results.** Comparison with prior work on different datasets. Top section between **red bars** shows *off-the-shelf supervised learning methods*; the middle section between **green bars** shows **self-supervised learning methods**; and the bottom section between **blue bars** shows **non-learning-based methods**. The off-the-shelf supervised learning methods were solely trained on synthetic datasets (models released by the authors). DCP was trained on ModelNet40; and FlowNet3D and PointPWC-Net were trained on FlyingThings3D. The self-supervised learning methods were trained on FlyingThings3D with supervision and then fine-tuned with self-supervision on each domain-specific dataset.  $\mathcal{E}$  is the mean absolute distance error, or end-point error;  $\%_5$  is the percentage of flow vectors where  $\mathcal{E} < 0.05$  or  $\delta < 5\%$ , where  $\delta$  is the percent error;  $\%_{10}$  is the percentage of flow vectors where  $\mathcal{E} < 0.1$  or  $\delta < 10\%$ ; and  $\theta_e$  is the mean angle error.  $\downarrow$  means smaller values are better and  $\uparrow$  means larger are better. All experiments were run with 2,048 points for each point cloud.

not generalize well to unseen data. ICP [4], in the non-learning section of Table 1, outperformed it.

FlowNet3D and PointPWC-Net predict the scene flow from pairwise point clouds. Both methods were trained with supervision on FlyingThings3D (*without fine-tuning*) and tested on all datasets. PointPWC-Net outperformed FlowNet3D which demonstrates that it better generalizes to unseen data. However, its performance is far from optimum given that supervised methods still struggle when tested on datasets that are statistically different from the training set.

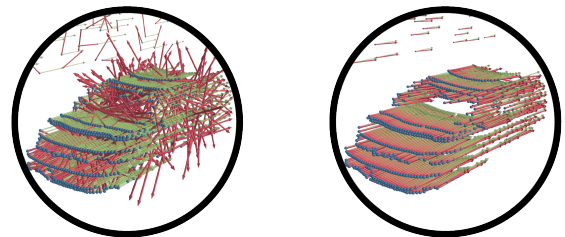
**With self-supervision.** We compared our method against two state-of-the-art self-supervised learning methods: Just Go with the Flow [23] and PointPWC-Net [45]. To the best of our knowledge, these are the only approaches that tackled self-supervision for scene flow estimation from point clouds. In this work, we are interested in the performance of the loss function by itself. However, we trained FlowNet3D on FlyingThings3D to use it as a baseline due to its simplicity. Then, we performed self-supervised fine-tuning on the KITTI, Argoverse, and nuScenes Scene Flow datasets using the loss functions proposed by Mittal *et al.* [23], Wu *et al.* [45], and ours. Our method outperformed the previous works showing that our proposed objective function can be successfully applied in self-supervised learning schemes.

### 5.2.2 Without learning

We compared our non-learning method against three non-learning methods: ICP, NICP, and PointPWC-Net (only the self-supervised loss was employed). Results are shown in the blue section of Table 1. We did not compare against Mittal *et al.* [23] since their self-supervised loss is network

	Without the graph Laplacian				With the graph Laplacian			
	$\mathcal{E} \downarrow$ (m)	$\%_5 \uparrow$ (%)	$\%_{10} \uparrow$ (%)	$\theta_e \downarrow$ (rad)	$\mathcal{E} \downarrow$ (m)	$\%_5 \uparrow$ (%)	$\%_{10} \uparrow$ (%)	$\theta_e \downarrow$ (rad)
• FlyingThings3D	0.636	1.58	5.83	1.059	<b>0.259</b>	<b>16.30</b>	<b>41.60</b>	<b>0.369</b>
• KITTI Scene Flow	0.882	1.59	4.12	1.123	<b>0.093</b>	<b>64.76</b>	<b>82.13</b>	<b>0.137</b>
• Argoverse Scene Flow	0.906	1.49	5.00	1.179	<b>0.257</b>	<b>25.26</b>	<b>47.50</b>	<b>0.467</b>
• nuScenes Scene Flow	1.080	1.84	5.11	1.181	<b>0.288</b>	<b>20.19</b>	<b>43.59</b>	<b>0.337</b>

Table 2. **Influence of the graph Laplacian.** Effect of the graph Laplacian regularizer on different datasets. Experiments were run without learning and each point cloud with 2,048 points.



(a) Without the graph Laplacian (b) With the graph Laplacian

Figure 3. **Influence of the graph Laplacian.** Qualitative effect of the graph Laplacian regularizer on a scene from Argoverse Scene Flow. The zoomed in region shows a point cloud that was sampled from a vehicle (blue points). Green and red arrows are the ground-truth and the predicted scene flow by our non-learning method, respectively. Note how chaotic the scene flow is if not regularized.

dependent. Our method achieved better performance in all metrics. An advantage of using our non-learning method is that it only needs two point clouds without annotations to estimate scene flow robustly. Also, since the objective is optimized during runtime, extra priors might be integrated as regularizes. Fig. 4 shows visual results. Scene flow were recovered with high fidelity for challenging scenes<sup>2</sup>.

<sup>2</sup>More examples in the supplementary material.

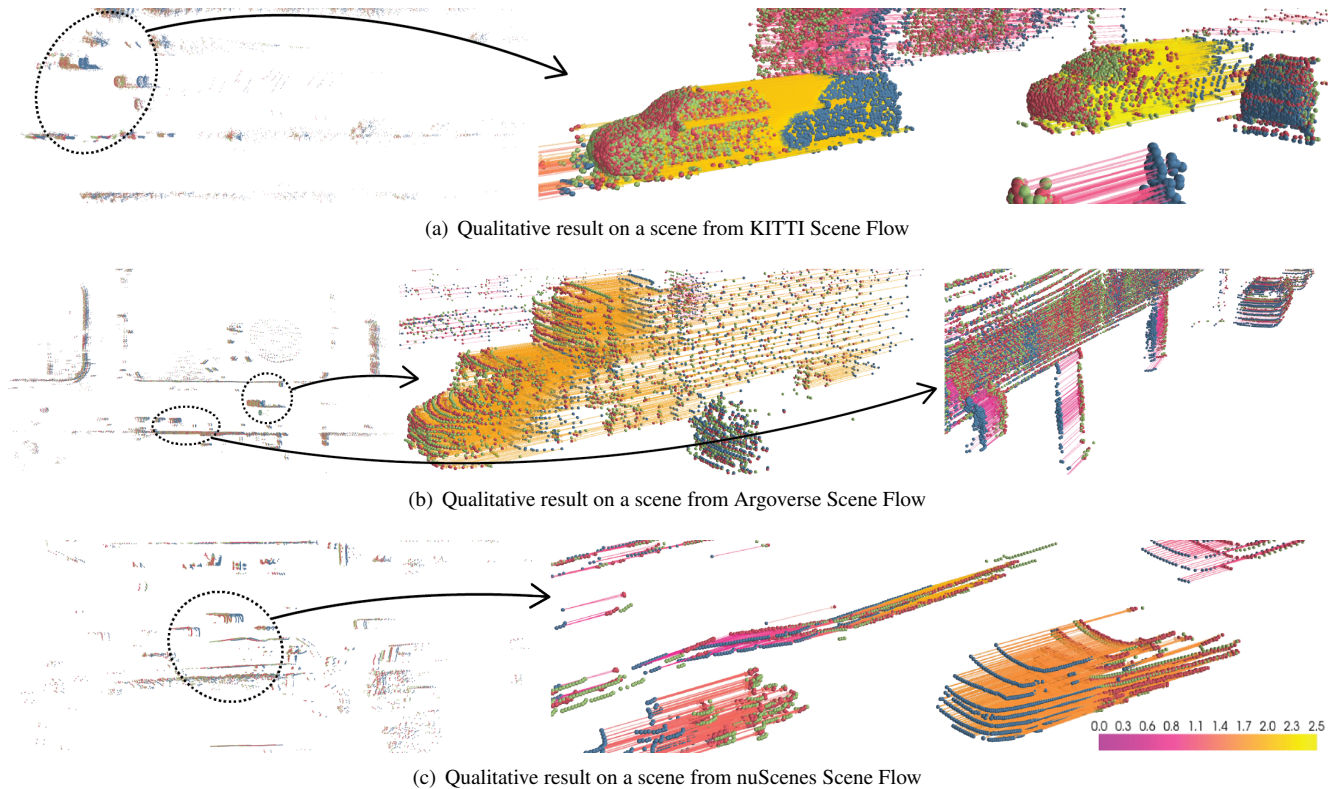


Figure 4. **Qualitative results.** Results of our non-learning method on dynamic scenes from (a) KITTI, (b) Argoverse, and (c) nuScenes Scene Flow. **Left:** Top view of the whole scene. **Right:** Zoomed in areas. **Blue** points are the source point cloud,  $\mathbf{P}_{t-1}$ , **green** points are the target point cloud,  $\mathbf{P}_t$ , **red** points are the translated source point cloud  $\mathbf{P}_{t-1} + \mathbf{F}$ . Arrows are the flows,  $\mathbf{F}$ , and its colors correspond to its magnitude (colorbar is shown in the bottom right). Note the different data modalities. For example, nuScenes has sparse measurements.

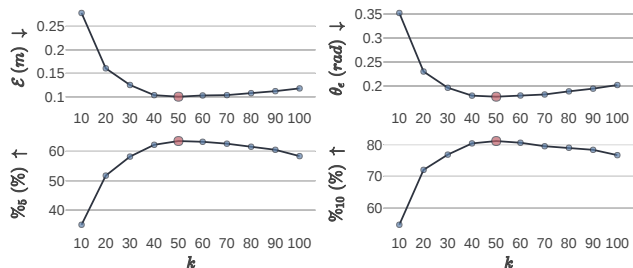


Figure 5. **Influence of the graph Laplacian.** Performance of our approach on KITTI Scene Flow when varying the number of neighbors,  $k$ , to form the graph. Number of points was fixed to 2,048 for each point cloud. Red marks are the maxima or minima.

### 5.2.3 Influence of the graph Laplacian

To understand the influence of the graph Laplacian regularizer, we removed it and evaluated the performance of the ablated model in Table 2. Visual results are shown in Fig. 3. The graph Laplacian regularizer drastically improved the performance in all metrics. We also tested the impact of the number of neighbors  $k$  to form the graph. Fig. 5 shows the results when fixing the number of points to 2,048 and varying  $k$ . Experiments were performed on KITTI. The best performance was with  $k=50$ . Larger  $k$ 's did not increase the performance but slightly hurt it. This is explained by the

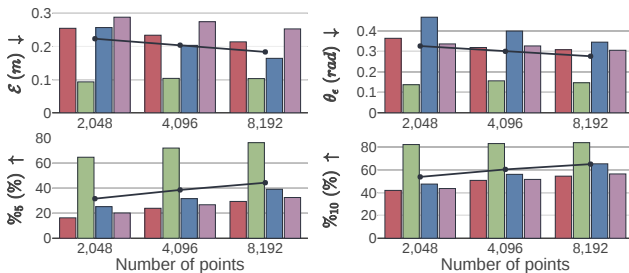


Figure 6. **Influence of the number of points.** Performance of our method when varying the number of points. **Red** FlyingThings3D; **Green** KITTI Scene Flow; **Blue** Argoverse Scene Flow; **Purple** nuScenes Scene Flow. Results improved with more points as shown in the trends.

fact that larger  $k$ 's encourages larger regions in the scene to move rigidly, *i.e.* the scene flow will be less flexible.

### 5.2.4 Influence of the number of points

We tested our non-learning model's performance when varying the number of points for both point clouds to 2,048, 4,096, and 8,192. In Fig. 6, we see that the performance grows as we increase the number of points. The denser the point clouds the easier it can be to search for better correspondences due to extra geometric information. However, a denser point cloud means a denser graph Laplacian due to

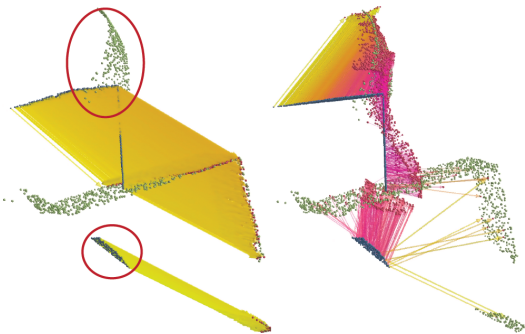


Figure 7. **Limitations.** Failure case in a FlyingThings3D scene. **Blue** is the source point cloud, **green** is the target point cloud, **red** is the translated source point cloud, and the arrows are the scene flow (colors correspond to its magnitude). **Left:** Shows the source, target, and ground-truth scene flow. **Right:** Shows the source, target, translated source, and the estimated scene flow by our method. The top red circle shows part of the target point cloud that is missing in the source point cloud. Thus the translated points got stuck in the wrong places due to bad correspondences. The bottom red circle shows part of the source point cloud that was wrongly translated due to the nearest-neighbors-based correspondence search.

more graph connectivity and a larger correspondence search space, thus increasing the computational complexity<sup>3</sup>. We also noted that the performance did not increase much for some metrics on the KITTI Scene Flow dataset when varying the number of points. This behavior suggests that the KITTI dataset preprocessed by [18] is close to performance saturation. Challenging datasets, such as FlyingThings3D, Argoverse, and nuScenes Scene Flow, are perhaps better suited to evaluate new scene flow methods.

### 5.3. Limitations and Discussions

We observed failure cases when the point clouds do not contain enough information to search for proper correspondences. This happens when the point clouds have large occlusions, holes, or missing parts, and these might lead to inaccurate scene flow estimation. Fig. 7 shows an example from FlyingThings3D where a change in visibility happened in the scene so that parts in the target point cloud is not visible in the source point cloud. Since we employed a nearest-neighbor-based method to search for correspondences, our approach will translate the source points to the nearest target points without knowledge of its local geometry and semantics. We also acknowledge that our proposed objective might not be optimal as there are several ways to construct the graph Laplacian. For example, one can use  $r$ -neighborhood graphs, “self-tuning” graphs, among others. However, we decided to formulate our method with the  $k$ -NN graph, which is simple and practical to use as a scene flow regularizer. Moreover, our non-learning method is orders of magnitude slower than our learning method (see runtime in 5.1). Nevertheless, it can be applied to cases that

<sup>3</sup>Please refer to the supplementary material for extra runtime analysis.



Figure 8. **Motion segmentation.** The red points have larger motions relative to the blue points. A truck, car, and pedestrian (bottom right) were segmented from the Argoverse Scene Flow scene.

require robustness instead of optimal speed. If robustness is not an issue, but rather time, our proposed objective function can be used to train a self-supervised model and act as a surrogate of our non-learning method.

### 5.4. Applications

**Motion segmentation.** Discontinuities in the scene flow can help to segment point clouds into regions that correspond to different objects. Fig. 8 shows an example of a scene segmentation. We simply set a threshold to filter large motions in the estimated scene flow.

**Point cloud densification.** Fig. 1 shows that our method can be applied to densify point clouds from dynamic scenes. Point cloud densification might be helpful to create dense depth maps from lidar point clouds and images. Five adjacent frames from an Argoverse scene in each direction were used to densify the current frame. Each point cloud has about 65e3 points. We visually compared our non-rigid densification against the original sparse point cloud and ICP. We did not use any semantic information nor a temporal consistency term across the frames for the registrations.

## 6. Conclusion

We presented a method to estimate the scene flow of dynamic scenes given a pair of point clouds. We proposed a simple and interpretable objective function to robustly approximate the scene flow with an “as-rigid-as-possible” regularizer based on the graph Laplacian. We successfully demonstrated that our objective function not only can be employed in self-supervised models when scene flow annotations are not available but also as a non-learning-based method in which the scene flow is optimized during runtime. Our approach outperformed the current self-supervised methods *with or without learning*.

**Acknowledgments.** We thank our Argo AI colleagues for their helpful suggestions. James Hays receives research funding from Argo AI, which is developing products related to the research described in this paper. In addition, the author serves as a principal scientist to Argo AI. The terms of this arrangement have been reviewed and approved by Georgia Tech in accordance with its conflict of interest policies.



## References

- [1] B. Amberg, S. Romdhani, and T. Vetter. Optimal step non-rigid ICP algorithms for surface registration. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.
- [2] T. Basha, Y. Moses, and N. Kiryati. Multi-view scene flow estimation: a view centered variational approach. *International Journal of Computer Vision (IJCV)*, 2013.
- [3] A. Behl, D. Paschalidou, S. Donne, and A. Geiger. PointFlowNet: learning representations for rigid motion estimation from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [4] P. J. Besl and N. D. McKay. A method for registration of 3D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 1992.
- [5] A. I. Bobenko and B. A. Springborn. A discrete Laplace—Beltrami operator for simplicial surfaces. *Discrete and Computational Geometry (DCG)*, 2007.
- [6] F. Brickwedde, S. Abraham, and R. Mester. Mono-SF: multi-view geometry meets single-view depth for monocular scene flow estimation of dynamic traffic scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019.
- [7] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuScenes: a multimodal dataset for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [8] J. Calder and N. G. Trillos. Improved spectral convergence rates for graph Laplacians on epsilon-graphs and k-NN graphs, 2019.
- [9] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, and J. Hays. Argoverse: 3D tracking and forecasting with rich maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [10] A. Dewan, T. Caselitz, G. D. Tipaldi, and W. Burgard. Rigid scene flow for 3D lidar scans. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, 2016.
- [11] M. Eisenberger, Z. Lahner, and D. Cremers. Smooth shells: multi-scale shape registration with functional maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [12] H. Fan, H. Su, and L. 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)*, 2017.
- [13] X. Gu, Y. Wang, C. Wu, Y. J. Lee, and P. 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)*, 2019.
- [14] F. Hugué and F. Devernay. A variational method for scene flow estimation from stereo sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.
- [15] J. Hur and S. Roth. Self-supervised monocular scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [16] D. P. Kingma and J. Ba. Adam: a method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.
- [17] F. Knöppel, K. Crane, U. Pinkall, and P. Schröder. Stripe patterns on surfaces. *ACM Transactions on Graphics (TOG)*, 2015.
- [18] X. Liu, C. R. Qi, and L. J. Guibas. FlowNet3D: learning scene flow in 3D point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [19] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. 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)*, 2016.
- [20] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [21] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [22] M. Menze, C. Heipke, and A. Geiger. *Joint 3D estimation of vehicles and scene flow*. Workshop on Image Sequence Analysis (ISA), 2015.
- [23] H. Mittal, B. Okorn, and D. Held. Just go with the flow: self-supervised scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [24] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: an imperative style, high-Performance deep learning library. In *Neural Information Processing Systems (NIPS)*, 2019.
- [25] J.-P. Pons, R. Keriven, and O. Faugeras. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. *International Journal of Computer Vision (IJCV)*, 2007.
- [26] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. PointNet++: deep hierarchical feature learning on point sets in a metric space. *Neural Information Processing Systems (NIPS)*, 2017.
- [27] N. Sharp, Y. Soliman, and K. Crane. The vector heat method. *Proceedings of SIGGRAPH*, 2019.
- [28] O. Sorkine. Laplacian mesh processing. *Eurographics (STARs)*, 2005.
- [29] O. Sorkine and M. Alexa. As-rigid-as-possible surface modeling. In *Proceedings of SIGGRAPH*, 2007.
- [30] L. Stankovic, D. P. Mandic, M. Dakovic, M. Brajovic, B. S. Dees, and T. Constantinides. Graph signal processing - part I: graphs, graph spectra, and spectral clustering. *CoRR*, abs/1907.03467, 2019.

- [31] P. Sun, H. Kretschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [32] F. Tombari, S. Salti, and L. D. Stefano. Unique signatures of histograms for local surface description. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010.
- [33] A. K. Ushani, R. W. Wolcott, J. M. Walls, and R. M. Eustice. A learning approach for real-time temporal scene flow estimation from lidar data. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2017.
- [34] L. Valgaerts, A. Bruhn, H. Zimmer, J. Weickert, C. Stoll, , and C. Theobalt. Joint estimation of motion, structure and geometry from stereo sequences. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010.
- [35] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1999.
- [36] C. Vogel, K. Schindler, and S. Roth. 3D scene flow estimation with a rigid motion prior. *International Journal of Computer Vision (IJCV)*, 2011.
- [37] C. Vogel, K. Schindler, and S. Roth. Piecewise rigid scene flow. *International Journal of Computer Vision (IJCV)*, 2013.
- [38] C. Vogel, K. Schindler, and S. Roth. 3D scene flow estimation with a piecewise rigid scene model. *International Journal of Computer Vision (IJCV)*, 2015.
- [39] C. Wang, B. Samari, and K. Siddiqi. Local spectral graph convolution for point set feature learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [40] Y. Wang and J. M. Solomon. Deep closest point: learning representations for point cloud registration. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019.
- [41] Z. Wang, S. Li, H. Howard-Jenkins, V. A. Prisacariu, and M. Chen. FlowNet3D++: geometric losses for deep scene flow estimation. In *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV)*, 2020.
- [42] A. Wedel, T. Brox, T. Vaudrey, C. Rabe, U. Franke, and D. Cremers. Stereoscopic scene flow computation for 3D motion understanding. *International Journal of Computer Vision (IJCV)*, 2011.
- [43] A. Wedel, C. Rabe, T. Vaudrey, T. Brox, U. Franke, , and D. Cremers. Efficient dense scene flow from sparse or dense stereo data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2008.
- [44] W. Wu, Z. Qi, and L. Fuxin. PointConv: deep convolutional networks on 3D point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [45] W. Wu, Z. Wang, Z. Li, W. Liu, and L. Fuxin. PointPWC-Net: a coarse-to-fine network for supervised and self-supervised scene flow estimation on 3D point clouds. In *arXiv:1911.12408v1*, 2019.
- [46] G. Yang and D. Ramanan. Upgrading optical flow to 3D scene flow through optical expansion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [47] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Neural Information Processing Systems (NIPS)*, 2004.
- [48] Q.-Y. Zhou, J. Park, and V. Koltun. Open3D: a modern library for 3D data processing. *arXiv:1801.09847*, 2018.

## Supplementary Material

### 1. Datasets

Here we provide additional information to supplement our main submission regarding the dataset preprocessing described in Section 5.1.

#### 1.1. Argoverse Scene Flow

Scene flow annotations are not provided in the Argoverse [9] dataset. To quantitatively evaluate our method, we created a dataset, “*Argoverse Scene Flow*”, based on the information provided in the Argoverse 3D Tracking v1.1 set. We used lidar point clouds sensed from two Velodyne VLP-32 sensors, six degrees of freedom (6DoF) vehicle poses, and the 3D object tracks to lift pseudo scene flow annotations. The 3D object tracks consist of bounding cuboids and poses for all objects of interest—both dynamic and static.

Given two consecutive lidar point clouds, we separated the rigid and non-rigid objects for both of them using the object track information. Then, we registered the rigid parts of both point clouds using the 6DoF vehicle pose, and the non-rigid segments were registered using the object poses. Thus the translational vectors can be recovered from the relative transformations to create the pseudo scene flow. Moreover, we used the ground height map information available in the Argoverse dataset to perform ground-points removal.

#### 1.2. nuScenes Scene Flow

Scene flow annotations are also not provided in the nuScenes [7] dataset. We used the same Argoverse preprocessing steps to create the “*nuScenes Scene Flow*”. However, since nuScenes does not provide ground height maps, we used RANSAC to remove the ground points<sup>4</sup>.

## 2. Qualitative Results

Here we provide additional qualitative results to Section 5.2 to supplement our main submission. Fig. 2–4 show visual results for our non-learning and self-supervised learning methods evaluated on the KITTI, Argoverse, and nuScenes Scene Flow datasets. The results show that our non-learning method provides scene flow estimation with higher fidelity than the self-supervised method.

## 3. Runtime

Here we provide new results to Section 5.2.4 to supplement our main submission. We tested our non-learning model’s performance when varying the number of points for both point clouds to 2,048, 4,096, and 8,192. Fig. 1 shows that the time grows as we increase the number of points. For this evaluation, we assumed that both point clouds, source

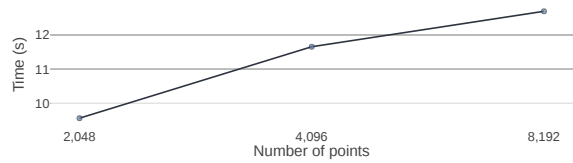
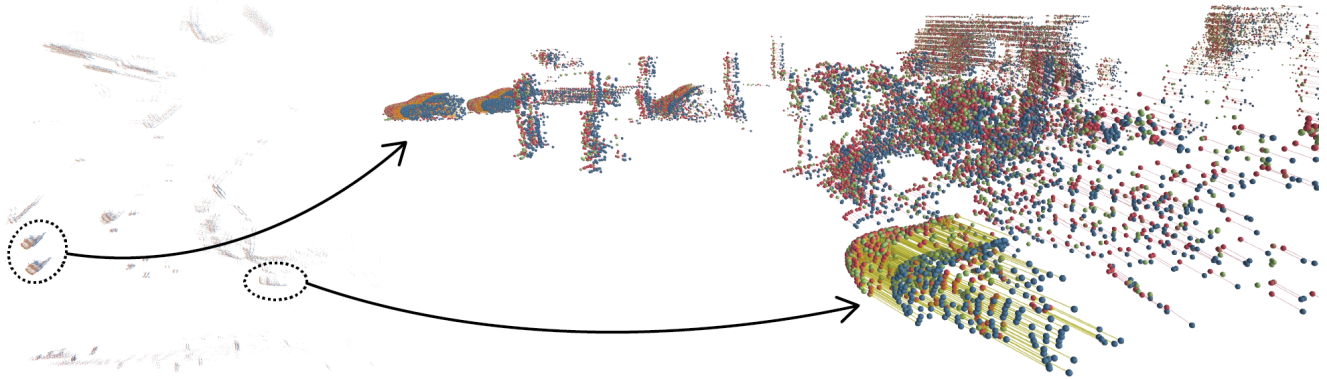


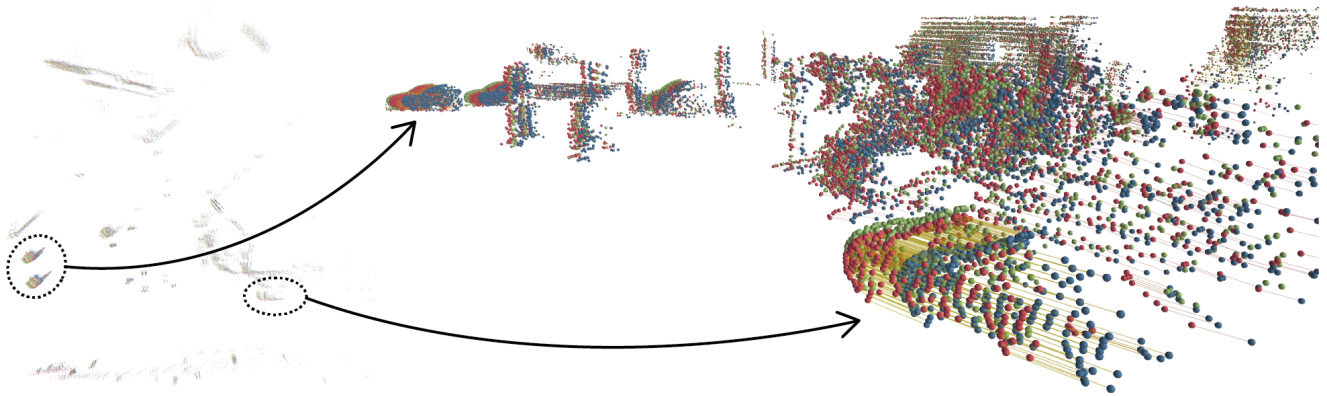
Figure 1. **Influence of the number of points.** Approximated runtime of our non-learning method. Time corresponds to 1,500 iterations that is needed for converging.

and target, have the same number of points. We tested the runtime on a single NVIDIA Quadro P5000 GPU.

<sup>4</sup>We used the Open3D library [48] to perform ground-point removal.

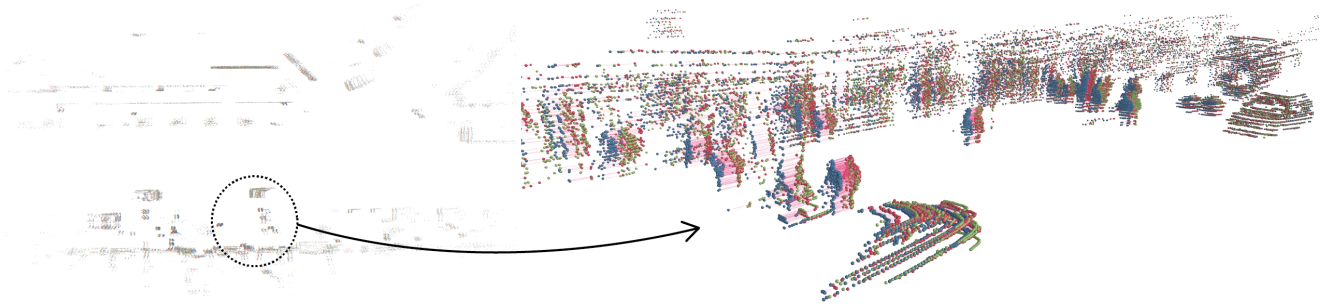


(a) Without learning: Qualitative result on a scene from KITTI Scene Flow

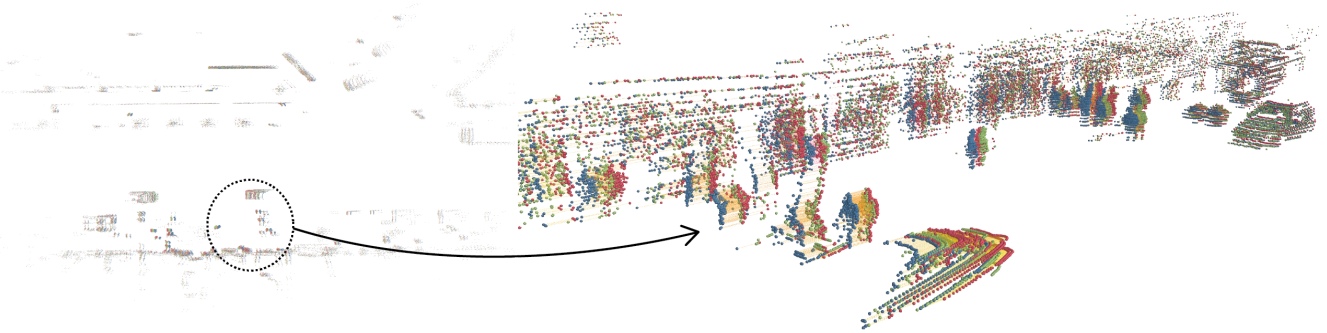


(b) With learning: Qualitative result on a scene from KITTI Scene Flow

Figure 2. **Qualitative results.** Results of our scene flow method on a dynamic scene from KITTI Scene Flow (a) without learning and (b) with self-supervised learning. **Left:** Top view of the whole scene. **Right:** Zoomed in areas. **Blue** points are the source point cloud,  $\mathbf{P}_{t-1}$ , **green** points are the target point cloud,  $\mathbf{P}_t$ , **red** points are the translated source point cloud  $\mathbf{P}_{t-1} + \mathbf{F}$ . Arrows are the flows,  $\mathbf{F}$ , and its colors correspond to its magnitude (yellowish is for bigger magnitudes). Note the better registration of our non-learning method.

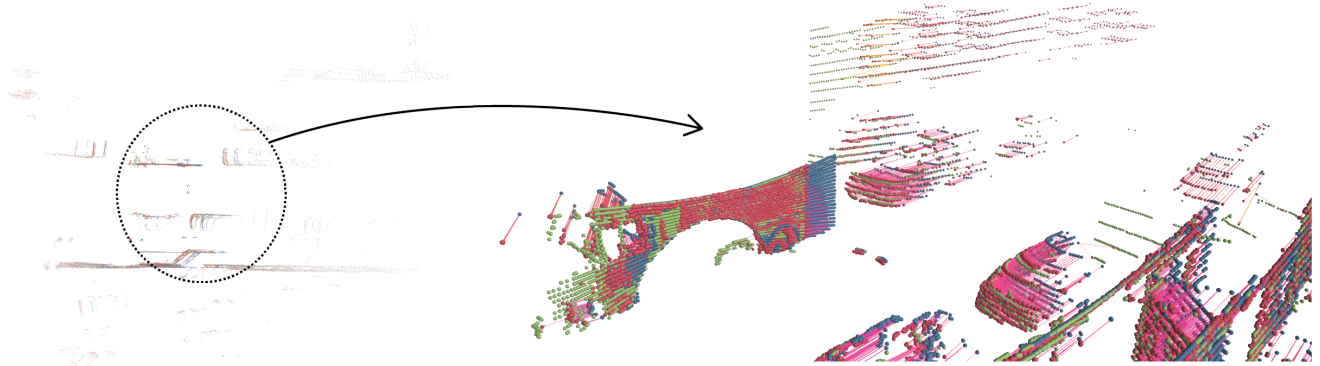


(a) Without learning: Qualitative result on a scene from Argoverse Scene Flow

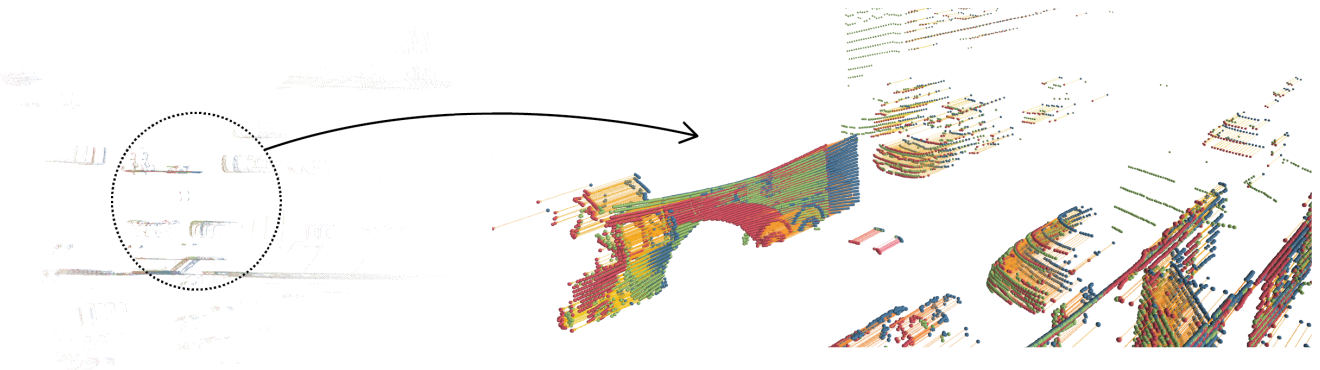


(b) With learning: Qualitative result on a scene from Argoverse Scene Flow

Figure 3. **Qualitative results.** Results of our scene flow method on a dynamic scene from Argoverse Scene Flow (a) without learning and (b) with self-supervised learning. **Left:** Top view of the whole scene. **Right:** Zoomed in areas. **Blue** points are the source point cloud,  $\mathbf{P}_{t-1}$ , **green** points are the target point cloud,  $\mathbf{P}_t$ , **red** points are the translated source point cloud  $\mathbf{P}_{t-1} + \mathbf{F}$ . Arrows are the flows,  $\mathbf{F}$ , and its colors correspond to its magnitude (yellowish is for bigger magnitudes). Note the better registration of our non-learning method.



(a) Without learning: Qualitative result on a scene from nuScenes Scene Flow



(b) With learning: Qualitative result on a scene from nuScenes Scene Flow

Figure 4. **Qualitative results.** Results of our scene flow method on a dynamic scene from nuScenes Scene Flow (a) without learning and (b) with self-supervised learning. **Left:** Top view of the whole scene. **Right:** Zoomed in areas. **Blue** points are the source point cloud,  $\mathbf{P}_{t-1}$ , **green** points are the target point cloud,  $\mathbf{P}_t$ , **red** points are the translated source point cloud  $\mathbf{P}_{t-1} + \mathbf{F}$ . Arrows are the flows,  $\mathbf{F}$ , and its colors correspond to its magnitude (yellowish is for bigger magnitudes). Note the better registration of our non-learning method, especially in the arch structure in the middle where a wrong scene flow was predicted by our self-supervised method.