

Just Go with the Flow: Self-Supervised Scene Flow Estimation

Himangi Mittal
Carnegie Mellon University
hmittal@andrew.cmu.edu

Brian Okorn
Carnegie Mellon University
bokorn@andrew.cmu.edu

David Held
Carnegie Mellon University
dheld@andrew.cmu.edu

Abstract

When interacting with highly dynamic environments, scene flow allows autonomous systems to reason about the non-rigid motion of multiple independent objects. This is of particular interest in the field of autonomous driving, in which many cars, people, bicycles, and other objects need to be accurately tracked. Current state-of-the-art methods require annotated scene flow data from autonomous driving scenes to train scene flow networks with supervised learning. As an alternative, we present a method of training scene flow that uses two self-supervised losses, based on nearest neighbors and cycle consistency. These self-supervised losses allow us to train our method on large unlabeled autonomous driving datasets; the resulting method matches current state-of-the-art supervised performance using no real world annotations and exceeds state-of-the-art performance when combining our self-supervised approach with supervised learning on a smaller labeled dataset.

1. Introduction

For an autonomous vehicle, understanding the dynamics of the surrounding environment is critical to ensure safe planning and navigation. It is essential for a self-driving vehicle to be able to perceive the actions of various entities around it, such as other vehicles, pedestrians, and cyclists. In the context of data recorded as 3D point clouds, a motion can be estimated for each 3D point; this is known as scene flow, which refers to the 3D velocity of each 3D point in a scene. Its 2D analog, optical flow, is the projection of scene flow onto the image plane of a camera. An alternative to scene flow estimation is to use 3D object detection for object-level tracking and to assume that all points within a bounding box have the same rigid motion. However, in such a pipeline, errors in object detection can lead to errors in tracking. In contrast, scene flow methods can avoid such errors by directly estimating the motion of each 3D point.

Recent state-of-the-art methods learn to estimate the scene flow from 3D point clouds [9, 6, 23, 26]. However,

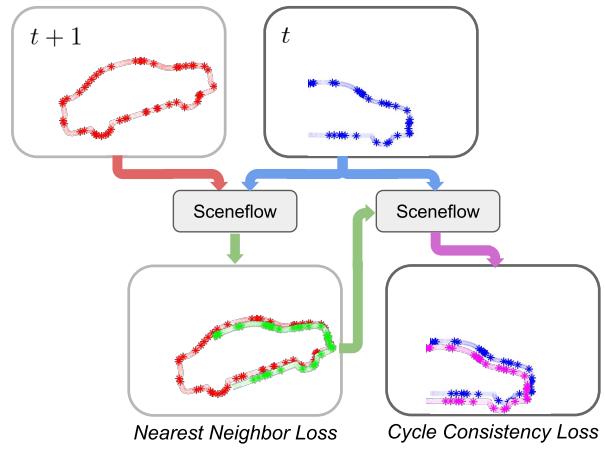


Figure 1: We use two self-supervised losses to learn scene flow on large unlabeled datasets. The “nearest neighbor loss” penalizes the distance between the predicted point cloud (green) and each predicted point’s nearest neighbor in the second point cloud (red). To avoid degenerate solutions, we also estimate the flow between these predicted points (green) in the reverse direction back to the original point cloud (blue) to form a cycle. The new predicted points from the cycle (purple) should align with the original points (blue); the distance between these two set of points forms our second self-supervised loss: “cycle consistency.”

these methods are fully supervised and require annotated datasets for training. Such annotations are costly to obtain as they require labeling the motion for every point in a scene. To compensate for the lack of real world data, learning-based methods for scene flow have been trained primarily on synthetic datasets and fine tuned on real world data. This requirement of labeled training data limits the effectiveness of such methods in real world settings.

To overcome this limitation, we propose a self-supervised method for scene flow estimation. Using a combination of two self-supervised losses, we are able to mimic the supervision created by human annotation. Specifically, we use a cycle consistency loss, which ensures that the scene flow produced is consistent in time (*i.e.* we ensure that a temporal cycle ends where it started). We also use

a nearest neighbor loss; due to the unavailability of scene flow annotations, we consider the nearest point to the predicted translated point, in the temporally next point cloud, as the pseudo-ground truth association. Intuitively, the nearest neighbor loss pushes one point cloud to flow toward occupied regions of the next point cloud. We show that this combination of losses can be used to train a scene flow network over large-scale, unannotated datasets containing sequential point cloud data. An overview of our method can be found in Figure 1.

We test our self-supervised training approach using the neural network architecture of a state-of-the-art scene flow method [9]. The self-supervision allows us to train this network on large-scale, unlabeled autonomous driving datasets. Our method matches the current state-of-the-art performance when no real world annotations are given. Moreover, our method exceeds the performance of state-of-the-art scene flow estimation methods when combined with supervised learning on a smaller labeled dataset.

2. Related Work

Scene Flow Vedula *et al.* [19] first introduced the task of scene flow estimation. They propose a linear algorithm to compute it from optical flow. Other works involve joint optimization of camera extrinsics and depth estimates for stereo scene flow [17], use of particle filters [7], local rigid motion priors [20, 22, 21, 11], and smoothness-based regularization [1].

Deep Scene Flow State-of-the-art scene flow methods today use deep learning to improve performance. FlowNet3D [9] builds on PointNet++ [16, 15] to estimate scene flow directly from a pair of point clouds. Gu *et al.* [6] produced similar results using a permutohedral lattice to encode the point cloud data in a sparse, structured manner. The above approaches compute scene flow directly from 3D point clouds [9, 6]. Methods involving voxelizations with object-centric rigid body assumptions [2], range images [18], and non-grid structured data [23] have also been used for scene flow estimation. All of the above methods were trained either with synthetic data [10] or with a small amount of annotated real-world data [5] (or both). Our self-supervised losses enable training on large unlabeled datasets, leading to large improvements in performance.

Self-Supervised Learning Wang *et al.* [25] used self-supervised learning for 2D tracking on video. They propose a tracker which takes a patch of an image at time t and the entire image at time $t - 1$ to track the image patch in the previous frame. They define a self-supervised loss by tracking the patch forward and backward in time to form a cycle while penalizing the errors through cycle consistency and feature similarity. We take inspiration from this work for our self-supervised flow estimation on point clouds. Other

works including self-supervisory signals are image frame ordering [14], feature similarity over time [24] for images, and clustering and reconstruction [8] from point clouds. While these can potentially be used for representation learning from 3D data, they cannot be directly used for scene flow estimation. Concurrent to our work, Wu *et al.* [26] showed that Chamfer distance, smoothness constraints, and Laplacian regularization can be used to train scene flow in a self-supervised manner.

3. Method

3.1. Problem Definition

For the task of scene flow estimation, we have a temporal sequence of point clouds: point cloud \mathcal{X} as the point cloud captured at time t and point cloud \mathcal{Y} captured at time $t + 1$. There is no structure enforced on these point clouds and they can be recorded directly from a LIDAR sensor or estimated through a stereo algorithm. Each point $p_i = \{x_i, f_i\}$ in point cloud \mathcal{X} contains the Cartesian position of the point, $x_i \in \mathbb{R}^3$, as well as any additional information which the sensor produces, such as color, intensity, normals, etc, represented by $f_i \in \mathbb{R}^c$.

The scene flow, $\mathcal{D} = \{d_i\}^N$, $d_i \in \mathbb{R}^3$ between these two point clouds describes the movement of each point x_i in point cloud \mathcal{X} to its corresponding position x'_i in the scene described by point cloud \mathcal{Y} , such that $x'_i = x_i + d_i$, and N is the size of point cloud \mathcal{X} . Scene flow is defined such that x_i and x'_i represent the same 3D point of an object moved in time. Unlike optical flow estimation, the exact 3D position of x'_i may not necessarily coincide with a point in the point cloud \mathcal{Y} , due to the sparsity of the point cloud. Additionally, the sizes of \mathcal{X} and \mathcal{Y} may be different.

Supervised Loss The true error associated with our task is the difference between the estimated flow $g(\mathcal{X}, \mathcal{Y}) = \hat{\mathcal{D}} = \{\hat{d}_i\}^N$ and the ground truth flow $\mathcal{D}^* = \{d_i^*\}^N$,

$$\mathcal{L}_{gt} = \frac{1}{N} \sum_i^N \|d_i^* - \hat{d}_i\|^2. \quad (1)$$

The loss in Equation 1 is useful because it is mathematically equivalent to the end point error, which we use as our evaluation metric. However, computing this loss requires annotated ground truth flow d_i^* . This type of annotation is easy to calculate in synthetic data [10], but requires expensive human labeling for real world datasets. As such, only a small amount of annotated scene flow datasets are available [12, 13]. While training on purely synthetic data is possible, large improvements can often be obtained by training on real data from the domain of the target application. For example, Lui *et al.* [9] showed an 18% relative improvement after fine-tuning on a small amount of annotated real world

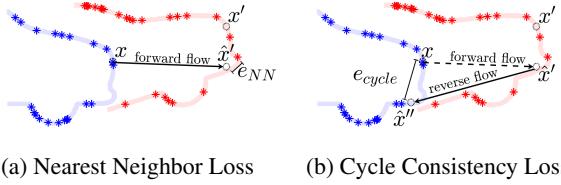


Figure 2: Example of our self-supervised losses between consecutive point clouds \mathcal{X} (blue) and \mathcal{Y} (red). We consider the point x whose ground truth projected point x' is not known during training time. (a) Nearest Neighbor Loss is computed between the projected point \hat{x}' , predicted by the forward flow, and the closest point in \mathcal{Y} . (b) The Cycle Consistency Loss tracks this transformed point back onto its original frame, as point \hat{x}'' , using the reverse flow, and computes the distance to its original position x .

data. This result motivates our work to use self-supervised training to train on large unlabeled datasets.

Nearest Neighbor (NN) Loss For large unlabeled datasets, since we do not have information about d_i^* , we cannot compute the loss in Equation 1. In lieu of annotated data, we take inspiration from methods such as Iterative Closest Point [3] and use the nearest neighbor of our transformed point $\hat{x}'_i = x_i + \hat{d}_i$ as an approximation for the true correspondence. For each transformed point $\hat{x}'_i \in \hat{\mathcal{X}}$, we find its nearest neighbor $y_j \in \mathcal{Y}$ and compute the Euclidean distance to that point, illustrated as e_{NN} in Figure 2a:

$$\mathcal{L}_{NN} = \frac{1}{N} \sum_i^N \min_{y_j \in \mathcal{Y}} \|\hat{x}'_i - y_j\|^2. \quad (2)$$

Assuming that the initial flow estimate is sufficiently close to the correct flow estimate, this loss will bring the transformed point cloud and the target point cloud closer. This loss can, however, have a few drawbacks if imposed alone. First, the true position of the point x_i transformed by the ground truth flow, $x'_i = x_i + d_i^*$, may not be the same as the position of the nearest neighbor to \hat{x}' (transformed by the estimated flow) due to potentially large errors in the estimated flow, as illustrated in Figure 2a. Further, the position of x'_i may not correspond with any point in \mathcal{Y} if the point cloud \mathcal{Y} is sufficiently sparse, as is common for point clouds collected by sparse 3D LIDAR for autonomous driving. Finally, this loss does not penalize degenerate solutions where all of the points in \mathcal{X} map to the same point in \mathcal{Y} ; such a degenerate solution would obtain 0 loss under Equation 2. To address these issues, we use an additional self-supervised loss: cycle consistency loss.

Cycle Consistency Loss To avoid the above issues, we incorporate an additional self-supervised loss: cycle consistency loss, illustrated in Figure 2b. We first estimate the “forward” flow as $\hat{\mathcal{D}} = g(\mathcal{X}, \mathcal{Y})$. Applying the estimated

flow $\hat{d}_i \in \hat{\mathcal{D}}$ to each point $x_i \in \mathcal{X}$ gives an estimate of the location of the point x_i in the next frame: $\hat{x}'_i = x_i + \hat{d}_i$. We then compute the scene flow in the reverse direction: for each transformed point \hat{x}'_i we estimate the flow to transform the point back to the original frame, $\hat{\mathcal{D}}' = g(\hat{\mathcal{X}}', \mathcal{X})$. Transforming each point \hat{x}'_i by this “reverse” flow \hat{d}'_i gives a new estimated point \hat{x}''_i . If both the forward and reverse flow are accurate, this point \hat{x}''_i should be the same as the original point x_i . The error between these points, e_{cycle} , is the “cycle consistency loss,” given by

$$\mathcal{L}_{cycle} = \sum_i^N \|\hat{x}''_i - x_i\|^2. \quad (3)$$

A similar loss is used as a regularization in [9].

However, we found that implementing the cycle loss in this way can produce unstable results if only self-supervised learning is used without any ground-truth annotations. These instabilities appear to be caused by errors in the estimated flow which lead to structural distortions in the transformed point cloud $\hat{\mathcal{X}}'$, which is used as the input for computing the reverse flow $g(\hat{\mathcal{X}}', \mathcal{X})$. This requires the network to simultaneously learn to correct any distortions in $\hat{\mathcal{X}}'$, while also learning to estimate the true reverse flow. To solve this problem, we use the nearest neighbor y_j of the transformed point \hat{x}'_i as an anchoring point in the reverse pass. Using the nearest neighbor y_j stabilizes the structure of the transformed cloud while still maintaining the correspondence around the cycle. The effects of this stabilization are illustrated in Figure 3. As we are using the anchoring point as part of the reverse pass of the cycle, we refer to this loss as “anchored cycle consistency loss”.

Specifically, we compute the anchored reverse flow as follows. First, we compute the forward flow as before, $\hat{\mathcal{D}} = g(\mathcal{X}, \mathcal{Y})$, which we use to compute the transformed point cloud $\hat{x}'_i = x_i + \hat{d}_i$. We then compute anchor points $\bar{\mathcal{X}}' = \{\bar{x}'_i\}^N$ as a convex combination of the transformed point and its nearest neighbor $\bar{x}'_i = \lambda \hat{x}'_i + (1 - \lambda)y_j$. In our experiments, we find that $\lambda = 0.5$ produces the most accurate results. Finally, we compute the reverse flow using these anchored points: $\bar{\mathcal{D}}' = g(\bar{\mathcal{X}}', \mathcal{X})$. The cycle loss of Equation 3 is then applied to this anchored reverse flow. By using anchoring, some of the structural distortion of the predicted point cloud $\hat{\mathcal{X}}'$ will be removed in the anchored point cloud $\bar{\mathcal{X}}'$, leading to a more stable training input for the reverse flow.

Note that the cycle consistency loss also has a degenerate solution: the “zero flow,” i.e. $\hat{\mathcal{D}} = 0$, will produce 0 loss according to Equation 3. However, the zero flow will produce a non-zero loss when anchored cycle consistency is used; thus anchoring helps to remove this degenerate solution. Further, the nearest neighbor loss will also be non-zero for the degenerate solution of zero flow.

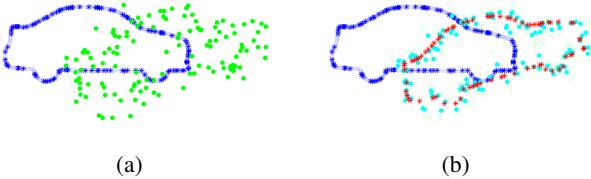


Figure 3: Compounding errors cause problems in estimating reverse flow using the transformed point cloud. (a) Large flow prediction errors degrade the structure of the transformed cloud $\hat{\mathcal{X}}'$ (shown in green). Thus, computing the reverse flow between $\hat{\mathcal{X}}'$ (green) and \mathcal{X} (blue) is an ill-posed task. (b) Using the nearest neighbor points (red) as anchors, we are able to stabilize the transformed cloud $\hat{\mathcal{X}}'$ (cyan), thus retaining important structural information.

Thus, the local minima of each of the nearest neighbor and cycle consistency losses conflict, allowing their sum, $\mathcal{L} = \mathcal{L}_{NN} + \mathcal{L}_{cycle}$, to act as a stable surrogate for the true error.

Temporal Flip Augmentation Having a dataset of point cloud sequences in only one direction may generate a motion bias which may lead to the network predicting the flow equal to the average forward speed of the training set. To reduce this bias, we augment the training set by flipping the point clouds, i.e, reversing the flow. With this augmentation, the network sees an equal number of point cloud sequences having a forward motion and a backward motion.

4. Experiments

We run several experiments to validate our self-supervised method for scene flow estimation for various levels of supervision and different amounts of data. First, we show that our method, with self-supervised training on large unlabeled datasets, can perform as well as supervised training on the existing labeled data. Next, we investigate how our results can be improved by combining self-supervised learning with a small amount of supervised learning, exceeding the performance of purely supervised learning. Finally, we explore the utility of each element of our method through an ablation study.

4.1. Implementation Details

For all data configurations (our method and the baseline), we initialize our network with the parameters of the Flownet3D model [9] pre-trained on the FlyingThing3D dataset [10]. We compare our self-supervised training procedure to a baseline which uses supervised fine-tuning on the KITTI dataset [5]. The baseline used in the comparison is the same as in Liu *et al.* [9], except that we increase the number of training iterations from 150 epochs (as described in the original paper) to 10k epochs in order to keep the number of training iterations consistent with that used in

our self-supervised method. We see that this change leads to a small improvement in the baseline performance, which we include in the results table.

4.2. Datasets

KITTI Vision Benchmark Suite KITTI [5] is a real-world self-driving dataset. There are 150 scenes of LIDAR data in KITTI collected using seven scans of a Velodyne 64 LIDAR, augmented using 3D models, and annotated with ground truth scene flow [13]. For our experiments under both self-supervised and supervised settings, we consider 100 out of 150 scenes for training and the remaining 50 scenes for testing. Ground points are removed from every scene using the pre-processing that was performed in previous work [9]. Every scene consists of a pair of point clouds recorded at two different times as well as the ground truth scene flow for every point of the first point cloud.

nuScenes The nuScenes [4] dataset is a large-scale public dataset for autonomous driving. It consists of 850 publicly available driving scenes in total from Boston and Singapore. The LIDAR data was collected using a Velodyne 32 LIDAR rotating at 20 Hz. This is in contrast to the 64-beam Velodyne rotating at 10 Hz used for the KITTI dataset. This difference in sensors leads to a difference in data sparsity that creates a distribution shift between KITTI and nuScenes. This distribution shift necessitates additional training on KITTI beyond our self-supervised training on nuScenes. Nonetheless, our results show a substantial benefit from the self-supervised training on nuScenes.

Since the nuScenes dataset [4] does not contain scene flow annotations, we must use self-supervised methods when working with this dataset. In our experiments, out of the 850 scenes available, we use 700 as the train set and the rest 150 as the validation set. Similar to KITTI, we remove the ground points from each point cloud using a manually tuned height threshold.

4.3. Results

We use three standard metrics to quantitatively evaluate the predicted scene flow when the ground truth annotations of scene flow are available. Our primary evaluation metric is End Point Error (EPE) which describes the mean Euclidean distance between the predicted and ground truth transformed points, described by Equation 1. We also compute accuracy at two threshold levels, $\text{Acc}(0.05)$ as the percentage of scene flow prediction with an $\text{EPE} < 0.05\text{m}$ or relative error $< 5\%$ and $\text{Acc}(0.1)$ as percentage of points having an $\text{EPE} < 0.1\text{m}$ or relative error $< 10\%$, as was done for evaluation in previous work [9].

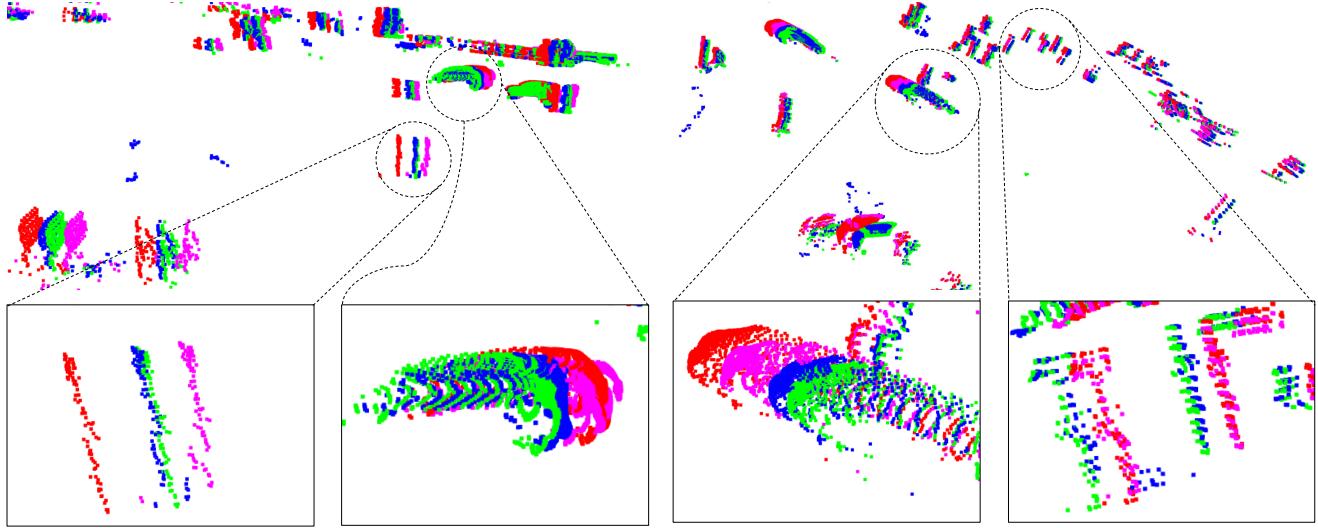


Figure 4: Scene flow estimation between point clouds at time t (red) and $t + 1$ (green) from the KITTI dataset trained without any labeled LIDAR data. Predictions from our self-supervised method, trained on nuScenes and fine-tuned on KITTI using self-supervised learning is shown in blue; the baseline with only synthetic training is shown in purple. In the absence of real-world supervised training, our method clearly outperforms the baseline method, which overestimate the flow in many regions. (Best viewed in color)

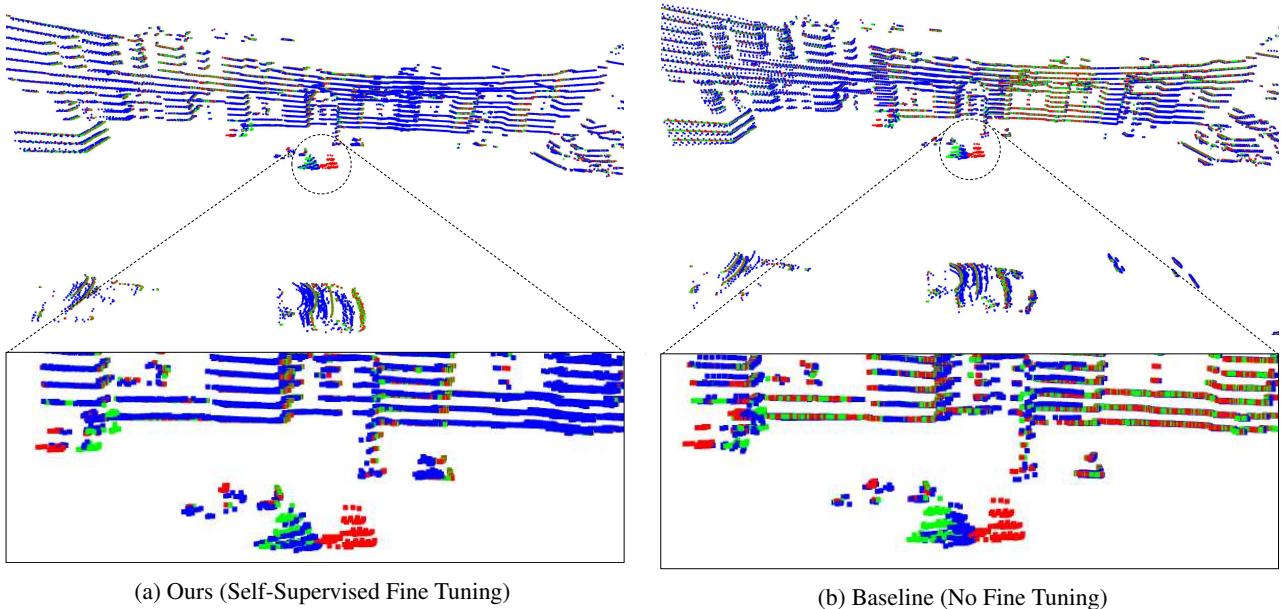


Figure 5: Comparison of our self-supervised method to a baseline trained only on synthetic data, shown on the nuScenes dataset. Scene flow is computed between point clouds at time t (red) and $t + 1$ (green); the point cloud that is transformed using the estimated flow is in shown in blue. In our method, the predicted point cloud has a much better overlap with the point cloud of the next timestamp (green) compared to the baseline. Since nuScenes dataset does not provide any scene flow annotation, the supervised approaches cannot be fine-tuned to this environment.

4.3.1 Quantitative Results

Self-supervised training: Unlike previous work, we are not restricted to annotated point cloud datasets; our method

can be trained on any sequential point cloud dataset. There are many point cloud datasets containing real LIDAR captures of urban scenes, but most of them do not contain scene flow annotations. Due to lack of annotations, these datasets

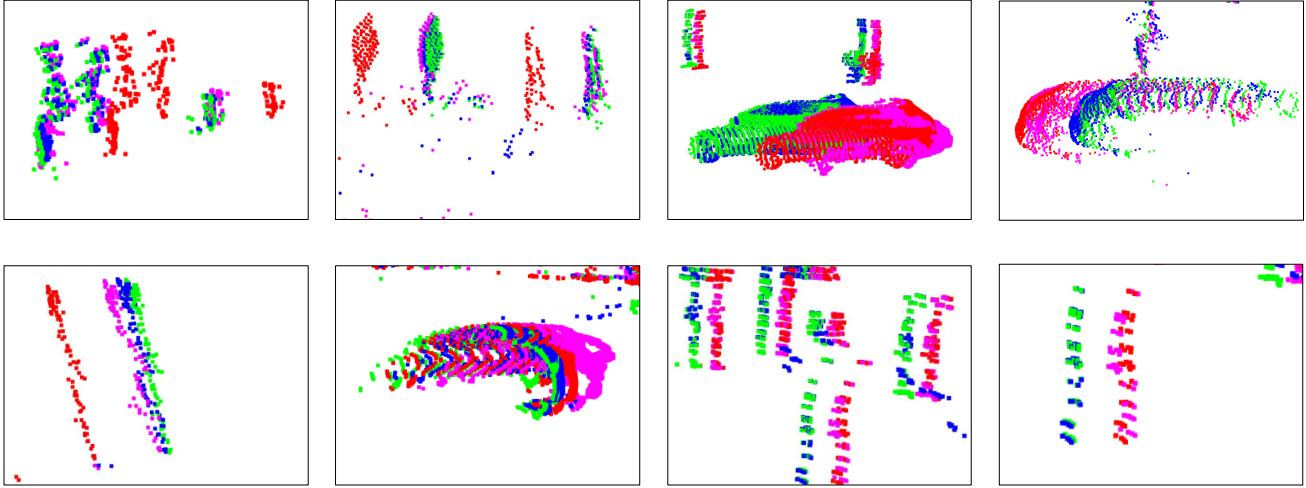


Figure 6: Scene flow estimation on LIDAR data from the KITTI dataset between point clouds at time t (red) and $t+1$ (green). Our method, which is trained on nuScenes using self-supervised learning and then fine-tuned on KITTI using supervised learning, is shown in blue. The baseline method is fine-tuned only on KITTI using supervised training and is shown in purple. While in aggregate, both methods well estimate the scene flow, adding self-supervised training on nuScenes (blue) enables our predictions to more closely match the next frame point cloud (green). In several scenes, the purely supervised method (purple) underestimates the flow, staying too close to the initial point cloud (red). (Best viewed in color)

| Training Method | EPE (m) \downarrow | ACC (0.05) \uparrow | ACC (0.1) \uparrow |
|--|----------------------|-----------------------|----------------------|
| No Fine Tuning | 0.122 | 25.37% | 57.85% |
| KITTI (Supervised) | 0.100 | 31.42% | 66.12% |
| Ablation: KITTI (Self-Supervised) | 0.126 | 32.00% | 73.64% |
| Ours: nuScenes (Self-Supervised) + KITTI (Self-Supervised) | 0.105 | 46.48% | 79.42% |
| Ours: nuScenes (Self-Supervised) + KITTI (Supervised) | 0.091 | 47.92% | 79.63% |

Table 1: Comparison of levels of supervision on KITTI dataset. The nearest neighbor + anchored cycle loss is used for nuScenes (self-supervised) and KITTI (self-supervised). All methods are pretrained on FlyingThings3D[10] and ground points are removed for KITTI and nuScenes datasets.

can not be utilized for supervised scene flow learning. In contrast, our self-supervised loss allows us to easily integrate them into our training set. The combination of these datasets (KITTI + NuScenes) contains 5x more real data than using KITTI alone.

The results are shown in Table 1. To show the value of self-supervised training, we evaluate the performance of our method without using any ground-truth annotations. We first pre-train on the synthetic FlyingThings3D dataset; we then perform self-supervised fine-tuning on the large nuScenes dataset followed by further self-supervised fine-tuning on the smaller KITTI dataset (4th row: “Ours: nuScenes (Self-Supervised) + KITTI (Self-Supervised)”). As can be seen, using no real-world annotations, we are able to achieve an EPE of 0.105 m. This outperforms the baseline of only training on synthetic data (“No Fine Tuning”). Even more impressively, our approach performs similarly to the baseline which pre-trains on synthetic data and then

does supervised fine-tuning on the KITTI dataset (“KITTI (Supervised)"); our method has a similar EPE and outperforms this baseline in terms of accuracy, despite not having access to any annotated training data. This result shows that our method for self-supervised training, with a large enough unlabeled dataset, can match the performance of supervised training.

Self-supervised + Supervised: Finally, we show the value of combining our self-supervised learning method with a small amount of supervised learning. For this analysis, we perform self-supervised training on NuScenes as above, followed by supervised training on the much smaller KITTI dataset. The results are shown in the last row of Table 1.

As can be seen, this approach of self-supervised training followed by supervised fine-tuning outperforms all other methods on this benchmark, obtaining an EPE of 0.091, outperforming the previous state-of-the-art result which used

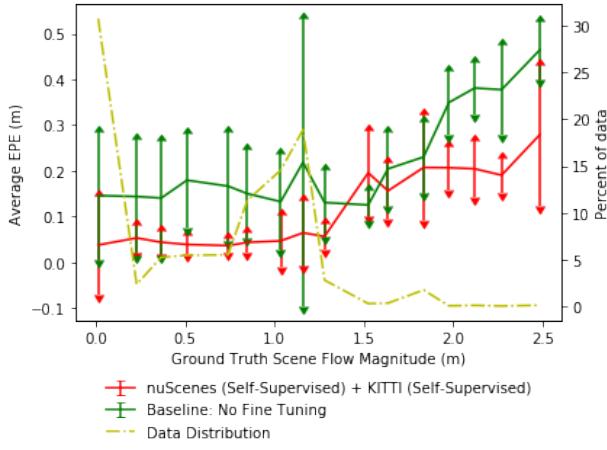


Figure 7: Analysis of average EPE (m) with respect to ground truth flow magnitudes (m). Flow estimates are binned by ground truth flow magnitude and a confidence interval of 95% is shown for all results.

only supervised training. This shows the benefit of self-supervised training on large unlabeled datasets to improve scene flow accuracy, even when scene flow annotations are available.

While Table 1 only shows results using the FlowNet3D [9] architecture, we note that our method also outperforms the results of HPLFlownet [6] (EPE of 0.1169) and all models they compare against as well.

Figure 7 provides an analysis on the correlation between average endpoint error and the magnitude of the ground truth flow. As can be seen, our method consistently outperforms the baseline at almost all flow magnitudes.

4.3.2 Qualitative Analysis

Self-supervised training - KITTI results: Next we perform a qualitative analysis to visualize the performance of our method. We compare our method (synthetic pre-training + self-supervised training on nuScenes + self-supervised training on KITTI) compared to the baseline of synthetic training only. Results on KITTI are shown in Figure 4. The figure shows the point clouds captured at time t and $t+1$ in red and green, respectively. The predictions from our method are shown in blue and the baseline predictions are shown in purple. As shown, our scene flow predictions (blue) have a large overlap with the point cloud at time $t+1$ (green). On the other hand, the baseline predictions (purple) do not overlap with the point cloud at time $t+1$. The baseline, trained only on synthetic data, fails to generalize to the real-world KITTI dataset. On the contrary, our self-supervised approach can be fine tuned on any real world environment and shows a significant improvement over the baseline.

| Self-Supervised Training (nuScenes + KITTI) | | | | | | |
|--|------------|------------|------|---------|------------|-----------|
| NN Loss | Cycle Loss | Anchor Pts | Flip | EPE (m) | ACC (0.05) | ACC (0.1) |
| ✓ | ✓ | ✓ | ✓ | 0.105 | 46.48% | 79.42% |
| ✓ | ✓ | ✓ | | 0.107 | 40.03% | 72.20% |
| ✓ | ✓ | | | 0.146 | 30.21% | 48.57% |
| ✓ | | | | 0.108 | 42.00% | 78.51% |
| | | | | 0.122 | 25.37% | 57.85% |
| Self-Supervised (nuScenes) + Supervised Training (KITTI) | | | | | | |
| NN Loss | Cycle Loss | Anchor Pts | Flip | EPE (m) | ACC (0.05) | ACC (0.1) |
| ✓ | ✓ | ✓ | ✓ | 0.091 | 47.92% | 79.63% |
| ✓ | ✓ | ✓ | | 0.093 | 40.69% | 74.50% |
| ✓ | ✓ | | | 0.092 | 30.76% | 72.94% |
| ✓ | | | | 0.114 | 31.24% | 64.58% |
| | | | | 0.100 | 31.42% | 66.12% |

Table 2: Ablation analysis: We study the effect of the different self-supervised losses and data augmentation. Top: Models use self-supervised training on nuScenes and KITTI; Bottom: Models use self-supervised training on nuScenes followed by supervised training on KITTI.

Self-supervised training - nuScenes results: Next, we visualize the performance of our method on the nuScenes dataset. Note that, because nuScenes does not have scene flow annotations, only qualitative results can be shown on this dataset. For this analysis, our method is pre-trained on synthetic data (FlyingThings3D) as before and then fine-tuned on nuScenes in a self-supervised manner. No scene flow annotations for nuScenes are available, so we compare to a baseline which is trained only on synthetic data.

The results on nuScenes are shown in Figure 5. These results again showcase the advantages of self-supervision on real world data over purely synthetic supervised training. As before, the figure shows the point clouds captured at time t and $t+1$ in red and green respectively. The predictions are shown in blue, with our method on the left (Fig. 5a) and the baseline on the right (Fig. 5b). As shown, our scene flow predictions (left figure, blue) have a large overlap with the point cloud at time $t+1$ (green). On the other hand, the baseline predictions (right figure, blue) do not overlap with the point cloud at time $t+1$.

The low performance of the baseline can again be attributed to its training on synthetic data and its inability to generalize to real-world data. For the nuScenes data, no scene flow annotations exist, so only self-supervised learning is feasible to improve performance.

Self-supervised + supervised training - KITTI results: Finally, we show the value of combining our self-supervised learning method with a small amount of supervised learning, compared to only performing supervised learning. For our method, we perform synthetic pre-training, followed by self-supervised fine-tuning on nuScenes, followed by supervised fine-tuning on the much smaller KITTI dataset. We compare to a baseline which only performs synthetic pre-training followed by supervised fine-tuning KITTI.

| λ | EPE (m) \downarrow | ACC (0.05) \uparrow | ACC (0.1) \uparrow |
|-----------|----------------------|-----------------------|----------------------|
| 0 | 0.120 | 24.09% | 73.20% |
| 0.25 | 0.122 | 26.41% | 65.57% |
| 0.5 | 0.105 | 46.48% | 79.42% |
| 0.75 | 0.125 | 23.59% | 62.96% |
| 1 | 0.149 | 22.97% | 49.58% |

Table 3: Effect of varying the λ parameter for “anchoring” the Cycle Consistency Loss. Results are shown for self-supervised training on nuScenes + KITTI.

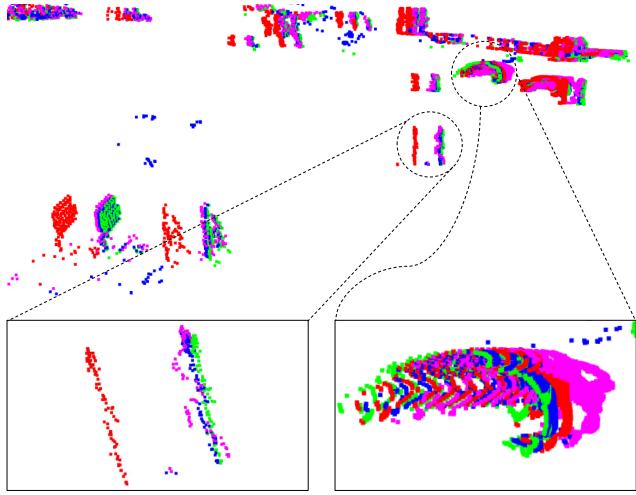


Figure 8: Ablation study comparing our self-supervised method with both nearest neighbor loss and anchored cycle consistency loss (blue) compared to training only using the nearest neighbor loss (purple). Scene flow is computed between point clouds from the KITTI dataset at time t (red) and $t + 1$ (green)

Qualitative results can be seen in Figure 6. The figure shows the point clouds captured at time t and $t + 1$ in red and green, respectively. The predictions from our method are shown in blue and the baseline predictions are shown in purple. As shown, our scene flow predictions (blue) have a large overlap with the point cloud at time $t + 1$ (green), whereas the baseline predictions (purple) do not.

The baseline predicts a small motion, keeping the transformed cloud (purple) too close to the initial position (red). As discussed above, this bias towards small motion is likely due to the training of the baseline over a synthetic dataset, which affects the generalization of the baseline to real-world datasets where objects exhibit different types of motion than seen in simulation. By training over a significantly larger unlabeled dataset, our method is able to avoid overfitting and generalizes better to the scenes and flow patterns which were not present in the synthetic dataset.

4.3.3 Ablation Studies

We test the importance of each component of our method by running a series of ablation studies. Table 2 shows the effect of iteratively removing portions of our method in both the purely self-supervised (top) and self-supervised + supervised (bottom) training. The benefits of anchored cycle consistency loss (compared to only using nearest neighbor loss) can be seen in Table 2 (bottom) as well as in Figure 8.

The benefits of anchoring is apparent by the large drop in performance for self-supervised training (top) when anchoring is removed. Introducing the anchor point cloud as part of the backward flow greatly improves performance when only self-supervised training is used (Figure 8, top). This suggests that having anchored point clouds stabilizes the training of the cycle consistency loss. Additionally, we evaluate the sensitivity of our system to the selection of the anchoring parameter, λ . Table 3 shows that we obtain the best results with $\lambda = 0.5$, *i.e.* using the average of the predicted and nearest points. Overall, these analyses show the benefits of each component of our method. Further ablation analysis can be found in the supplement.

5. Conclusion

In this work, we propose a self-supervised method for training scene flow algorithms using a combination of cycle consistency in time and nearest neighbor losses. Our purely self-supervised method is able to achieve a performance comparable to that of the supervised methods on the widely used KITTI self-driving dataset. We further show that when supervised training is augmented with self-supervision on a large-scale, unannotated dataset, the results exceed the current state-of-the-art performance. Our self-supervision method opens the door to fine-tuning on arbitrary datasets that lack scene flow annotations.

Acknowledgements

This work was supported by the CMU Argo AI Center for Autonomous Vehicle Research and a NASA Space Technology Research Fellowship.

References

- [1] Tali Basha, Yael Moses, and Nahum Kiryati. Multi-view scene flow estimation: A view centered variational approach. *IJCV*, 101(1), 2013. 2
- [2] Aseem Behl, Despoina Paschalidou, Simon Donne, and Andreas Geiger. Pointflownet: Learning representations for rigid motion estimation from point clouds. In *CVPR*, 2019. 2
- [3] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. International Society for Optics and Photonics, 1992. 3

- [4] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liang, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019. 4
- [5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 2, 4
- [6] Xiuye Gu, Yijie Wang, Chongruo Wu, Yong Jae Lee, and Panqu Wang. Hplflownet: Hierarchical permutohedral lattice flownet for scene flow estimation on large-scale point clouds. In *CVPR*, 2019. 1, 2, 7
- [7] Simon Hadfield and Richard Bowden. Kinecting the dots: Particle based scene flow from depth sensors. In *ICCV*, 2011. 2
- [8] Kaveh Hassani and Mike Haley. Unsupervised multi-task feature learning on point clouds. In *ICCV*, 2019. 2
- [9] Xingyu Liu, Charles R Qi, and Leonidas J Guibas. Flownet3d: Learning scene flow in 3d point clouds. In *CVPR*, 2019. 1, 2, 3, 4, 7
- [10] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016. 2, 4, 6
- [11] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015. 2
- [12] Moritz Menze, Christian Heipke, and Andreas Geiger. Joint 3d estimation of vehicles and scene flow. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2, 2015. 2
- [13] Moritz Menze, Christian Heipke, and Andreas Geiger. Object scene flow. *ISPRS Journal of Photogrammetry and Remote Sensing*, 140, 2018. 2, 4
- [14] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*, 2016. 2
- [15] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017. 2
- [16] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099–5108, 2017. 2
- [17] Levi Valgaerts, Andrés Bruhn, Henning Zimmer, Joachim Weickert, Carsten Stoll, and Christian Theobalt. Joint estimation of motion, structure and geometry from stereo sequences. In *ECCV*, 2010. 2
- [18] Victor Vaquero, Alberto Sanfeliu, and Francesc Moreno-Noguer. Deep lidar cnn to understand the dynamics of moving vehicles. In *ICRA*, 2018. 2
- [19] Sundar Vedula, Simon Baker, Peter Rander, Robert Collins, and Takeo Kanade. Three-dimensional scene flow. In *ICCV*, 1999. 2
- [20] Christoph Vogel, Konrad Schindler, and Stefan Roth. 3d scene flow estimation with a rigid motion prior. In *ICCV*, 2011. 2
- [21] Christoph Vogel, Konrad Schindler, and Stefan Roth. Piecewise rigid scene flow. In *ICCV*, 2013. 2
- [22] Christoph Vogel, Konrad Schindler, and Stefan Roth. 3d scene flow estimation with a piecewise rigid scene model. *IJCV*, 115(1), 2015. 2
- [23] Shenlong Wang, Simon Suo, Wei-Chiu Ma, Andrei Pokrovsky, and Raquel Urtasun. Deep parametric continuous convolutional neural networks. In *CVPR*, 2018. 1, 2
- [24] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, 2015. 2
- [25] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *CVPR*, 2019. 2
- [26] Wenzuan Wu, Zhiyuan Wang, Zhuwen Li, Wei Liu, and Li Fuxin. Pointpwc-net: A coarse-to-fine network for supervised and self-supervised scene flow estimation on 3d point clouds. 2019. 1, 2