4D Forecasting: Sequential Forecasting of 100,000 Points

Xinshuo Weng[1], Jianren Wang[1], Sergey Levine[2], Kris Kitani[1], and Nicholas Rhinehart[2]

[1] Robotics Institute, Carnegie Mellon University
{xinshuow, jianrenw, kkitani}@cs.cmu.edu
[2] Berkeley Artificial Intelligence Research Lab, University of California, Berkeley
{svlevine, nrhinehart}@eecs.berkeley.edu

Abstract. Predicting the future is a crucial first step to effective control. In this work, we study the problem of future prediction of 3D scenes, represented by point clouds captured by a LiDAR sensor, i.e., directly forecasting the evolution of >100,000 points that comprise a complete scene. We term this Sequential Pointcloud Forecasting (SPF). By directly predicting the densest-possible 3D representation of the future, the output contains richer information than output of prior forecasting tasks such as future object trajectories. We design a method, SPFNet, evaluate it on the KITTI and nuScenes datasets, and find that it demonstrates excellent performance on the SPF task. Our project website is at http://www.xinshuoweng.com/projects/SPF2.

Keywords: 3D point cloud, forecasting

1 Introduction

Forecasting the future is crucial in applications such as autonomous driving [10], as the ability to forecast is often the first step toward planning and control. Prior forecasting tasks such as trajectory forecasting [6,14] and activity forecasting [8], forecast at the object level. Specifically, these tasks only predict the future pose and action categories of the agents, rather than the future of everything in the scene. To train learning-based models for object-level forecasting tasks, object pose or action labels are typically required, which are costly to obtain, especially in 3D space [3], and may not be accurate.

To circumvent this labeling bottleneck and enable forecasting with rich scene information, we seek a forecasting task that (1) predicts the densest-possible 3D representation (i.e. a point cloud) of the future, containing rich information about the scene and objects; (2) requires no human annotation in order to train at a large scale, e.g., ground truth can be accurate observations captured by a LiDAR sensor. We term this task Sequential Pointcloud Forecasting (SPF), which is to forecast a sequence of point clouds for an entire scene given past point clouds. We illustrate the difference between object trajectory forecasting task and the SPF task in Fig. 1. Our proposed SPF can be viewed as the 3D counterpart of the video forecasting task, although it requires careful treatment of unordered 3D point clouds rather than fixed-size 2D images.

To solve the proposed SPF task, we present the SPFNet, which employs an LSTM autoencoder model and is end-to-end trainable. To leverage the 2D Convolutional Neural Networks (CNNs) while preserving 3D structure of the LiDAR

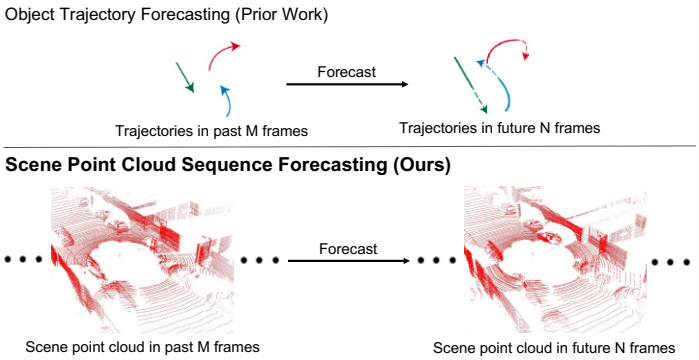

Fig. 1: (Top) object trajectory forecasting: given the object trajectories in past M frames, the goal is to predict the object trajectories in future N frames. **(Bottom) Sequential Pointcloud Forecasting**: given 3D point clouds for an entire scene in past M frames, the goal is to predict a sequence of future scene point clouds.

point cloud, we use the range map representation [2] in our SPFNet. We use the raw LiDAR point cloud data from the KITTI [5] and nuScenes [3] datasets for training and evaluation, showing that our SPFNet can reliably forecast the future point clouds upto three seconds and outperform competitive baselines that we devised from existing S.O.T.A. techniques. Note that our proposed SPFNet can be further utilized for standard trajectory forecasting task by applying an off-the-shelf 3D detector [12] and tracker [13] on the generated future point clouds by our SPFNet. Our contributions are summarized as follows:

1. **A new task, Sequential Pointcloud Forecasting**, which predicts the densest-possible future and does not require any human annotation;
2. **An effective approach for SPF**, deemed SPFNet, that outperforms competitive approaches we devised from S.O.T.A. techniques.

2 Task: Sequential Pointcloud Forecasting

The goal of our SPF is to predict a sequence of future scene point cloud given a sequence of past scene point clouds. Specifically, given M frames of past scene point clouds with each frame $S_t = \{(x, y, z)_j\}_{j=1}^{K_t}$, where $t \in [-M+1, \cdots, 0]$ denotes the frame index, $j \in [1, \cdots, K_t]$ denotes the index of points and K_t denotes the number of points at frame t, the goal of SPF is to predict N frames of future scene point clouds with each frame $S_t = \{(x, y, z)_j\}_{j=1}^{K_t}$, where $t \in [1, \cdots, N]$. Note that the number of points K_t can be different across frames (e.g., the point cloud captured by a standard Velodyne LiDAR sensor often has different number of points across frames).

3 Approach: SPFNet

Our proposed SPFNet is shown in Fig. 2, which consists of four modules: (a) a shared encoder that first converts the point cloud to a 2D range map and then uses CNNs to extract features, (b) an LSTM for temporal modeling, (c) a shared decoder that uses deconvolutional neural networks to convert the LSTM output features to a range map and then reshape it into a 3D point cloud, (d) three

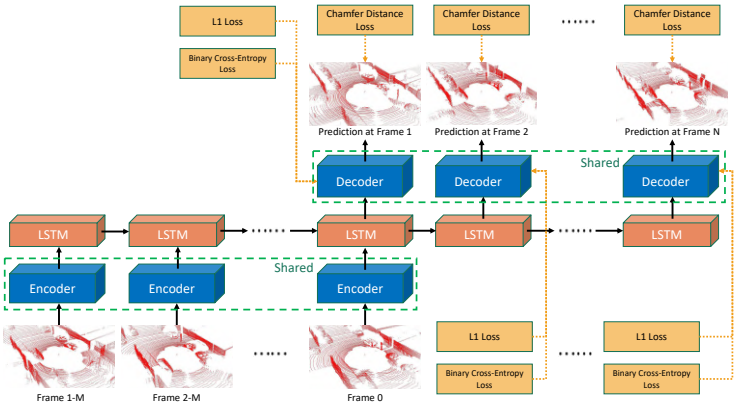

Fig. 2: Proposed SPFNet.

losses: a Chamfer distance loss applied between the ground truth and the final generated point cloud, and L1 loss + binary cross entropy loss applied between the ground truth range map and our range map estimated from the decoder.

4 Experiments

Dataset and Evaluation Metrics. We evaluate on the KITTI raw [5] and nuScenes [3], each with 156 and 1000 point cloud sequences. For KITTI, we split the train, val, and test each with 126/15/15 sequences. For nuScenes, we use the default splits. We use Earth Mover's Distance (EMD) [9] and Chamfer Distance (CD) [4] to measure the distance between two point clouds. We evaluate two settings for all methods: (1) observe 1.0 second past and predict the next 1.0 second; (2) observe 3.0 seconds past and predict the next 3.0 seconds.

Baselines. As we are the first towards solving SPF, there is no direct baseline to compare. Therefore, we devise competitive baselines based on existing techniques: (1) *Identity*. We duplicate the point cloud from the last frame to future for N times; (2) *GT-Ego*. We use the ground truth ego-motion (rotation and translation) between adjacent past frames and compute an average motion, which is used to warp the last frame of point cloud to future frames; (3) *Est-Ego*. Instead of using ground truth ego-motion, we estimate the ego-motion using [15] for warping; (4) *Align*. We use point cloud alignment methods (ICP [1] and Deep-ICP [11]) to compute the global rigid motion between adjacent past frames and obtain the average motion for warping; (5) *SceneFlow*. We use scene flow methods [7] to estimate point-wise motion between the last two frames and use the motion to warp the point cloud to future frames. For baselines that require training, we fine-tune their models on KITTI and nuScenes datasets.

Results. We summarize the results in Table 1. The proposed SPFNet consistently outperforms all competitive baselines in both EMD and CD metrics. We believe it is because our SPFNet (1) is designed for sequence prediction and (2) can learn the future location for each individual point. On the other hand, the baselines are originally designed for one-frame prediction (*i.e.*, warping between two frames), leading to poor performance in long-horizon prediction. Also,

Table 1: Evaluation for the proposed SPF task on KITTI and nuScenes datasets.

Datasets	Metrics	Identity	GT-Ego	Est-Ego	Align-ICP	Align-[11]	SceneFlow	**Ours**
KITTI-1.0s	CD↓	12.82	5.47	9.18	6.13	6.02	3.15	0.89
	EMD↓	526.87	391.03	495.21	418.25	439.17	291.63	128.81
KITTI-3.0s	CD↓	13.31	7.91	11.31	9.14	9.57	5.08	0.94
	EMD↓	602.89	452.81	502.83	470.25	493.26	351.46	175.54
nuScenes-1.0s	CD↓	8.42	2.16	4.91	4.04	3.50	1.93	0.35
	EMD↓	461.63	168.37	299.13	281.53	270.81	117.41	78.37
nuScenes-3.0s	CD↓	10.16	2.85	6.52	7.13	5.27	3.81	0.41
	EMD↓	494.81	190.14	370.91	419.37	332.97	294.53	91.83

most baselines can only estimate the global rigid motion and do not account for point-wise motion except for [7].

5 Conclusion

We proposed a new forecasting task, Sequential Pointcloud Forecasting, and a method SPFNet, which predicts the densest-possible 3D representation of the future and does not need any human annotation for training.

References

1. Besl, P., McKay, N.: A Method for Registration of 3-D Shapes. TPAMI (1992)
2. Caccia, L., Herke Van Hoof, Courville, A., Pineau, J.: Deep Generative Modeling of LiDAR Data. IROS (2019)
3. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q.: nuScenes: A Multimodal Dataset for Autonomous Driving. CVPR (2020)
4. Fan, H., Su, H., Guibas, L.: A Point Set Generation Network for 3D Object Reconstruction from a Single Image. CVPR (2017)
5. Geiger, A., Lenz, P., Urtasun, R.: Are We Ready for Autonomous Driving? the KITTI Vision Benchmark Suite. CVPR (2012)
6. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks. CVPR (2018)
7. Liu, X., Qi, C.R., Guibas, L.J.: FlowNet3D: Learning Scene Flow in 3D Point Clouds. CVPR (2019)
8. Rhinehart, N., Kitani, K.M.: First-Person Activity Forecasting from Video with Online Inverse Reinforcement Learning. TPAMI (2018)
9. Rubner, Y., Tomasi, C., Guibas, L.J.: The Earth Mover's Distance as a Metric for Image Retrieval. IJCV (2000)
10. Wang, S., Jia, D., Weng, X.: Deep Reinforcement Learning for Autonomous Driving. arXiv:1811.11329 (2018)
11. Wang, Y., Solomon, J.M.: Deep Closest Point: Learning Representations for Point Cloud Registration. ICCV (2019)
12. Weng, X., Kitani, K.: Monocular 3D Object Detection with Pseudo-LiDAR Point Cloud. ICCVW (2019)
13. Weng, X., Wang, J., Held, D., Kitani, K.: 3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. IROS (2020)
14. Weng, X., Yuan, Y., Kitani, K.: Joint 3D Tracking and Forecasting with Graph Neural Network and Diversity Sampling. arXiv:2003.07847 (2020)
15. Zhou, T., Brown, M., Noah, G., Google, S., Lowe Google, D.G.: Unsupervised Learning of Depth and Ego-Motion from Video. CVPR (2017)