Re-thinking LiDAR-Stereo Fusion Frameworks

Qilin Jin
Department of Computer Science
University of North Carolina
jinql@cs.unc.edu

Parasara Sridhar Duggirala
Department of Computer Science
University of North Carolina
psd@cs.unc.edu

Abstract

In this paper, we present a 2-step framework for high-precision dense depth perception from stereo RGB images and sparse LiDAR input. In the first step, we train a deep neural network to predict dense depth map from the left image and sparse LiDAR data, in a novel self-supervised manner. Then in the second step, we compute a disparity map from the predicted depths, and refining the disparity map by making sure that for every pixel in the left, its match in the right image, according to the final disparity, is the local optimum.

Introduction

3D information perception has been playing an essential role in numerous robotics and computer vision tasks, such as autonomous driving, 3D object detection, and 3D reconstruction. To obtain reliable dense depth information from the scene, stereo cameras (with stereo matching algorithms), and active sensing sensors (e.g. 3D LiDARs) are the most commonly used techniques. While each of them alone has inherent pros and cons, stereo cameras can provide two rectified RGB images with dense information, but the necessary stereo matching algorithms, which computes the disparity for each pixel in the left image by finding its corresponding pixel in the right image, are well-known to be expensive in computation, and problematic in matching regions with repetitive patterns, homogeneous appearance, and occluded objects. Alternatively, 3D LiDAR scanners can measure the depths accurately, but their outputs are usually too sparse, and dense 3D LiDARs are expensive. Thus, fusing sparse 3D LiDAR data with images becomes a promising option in computing a trustworthy dense depth map.

Previously, the general idea for this task is to plug LiDAR data into existing stereo matching algorithms. For example, Wang et al (2019) enhance the GC-Net (Kendall et al, 2017) by extracting more features and regularizing the cost volume using LiDAR data, while Park et al (2018) design a deep learning network to refine the output of SGM (Hirschmüller, 2008) with the help of LiDAR data. However, we argue that these stereo matching-base algorithms have two intrinsic drawbacks: first, intuitively, once the accurate disparity value of a pixel is known, it should usually be easier to compute the disparities of the pixels around it, thus faster to compute a dense disparity map. However, none of the state-of-art methods actually reduce the computational burden of their corresponding stereo matching algorithms, instead they exert more; secondly, even though deep learning methods are achieving much better results than others, due to its lack of explainability, it would be hard to ensure the safety of their outputs.

Thus, to make LiDAR-stereo fusion more efficient and explainable, and to improve its accuracy, we propose a brand new 2-step framework for LiDAR-Stereo fusion problems: in the first step, we take advantage of the progress in deep learning and use a convolutional neural network to predict a raw depth map from the left image and sparse LiDAR data; next, in the second step, instead of searching the whole disparity range to find the best match, we give confidence to the disparity transformed from the predicted depth map, and refine it by only looking for a better match around it.

Framework Details

Step 1: Disparity Prediction

The architecture design of our LiDAR-Monocular fusion network, which follows an encoder-decoder paradigm, is based on the work of Ma et al (2019). The biggest modification is, we train the network in a different self-supervised way, and reach the same level of accuracy as the state-of-art. As shown in the figure 1, during the training process, besides normal depth loss computed by summing up the differences between raw prediction and the ground truth, we also introduce a new kind of photometric loss, which represents the difference between the right image and the warped left image, to the loss function. Specifically, if we use $I_L(i,j)$ to represent the the value of the pixel at the i_{th} row, j_{th} column of the left image, and its predicted disparity is $D(i,j)$, then, the value of its corresponding pixel in the warped image $I_{L_{warp}}$ will be:

$$I_{L_{warp}}(i, j + D(i,j)) = I_L(i,j).$$

In this way, if we the raw disparity prediction is generally correct, the warped image $I_{L_{warp}}$ should look like the right

image, with some points being zero if no pixel in the left is mapped to that location. Thus, we define our photometric loss to be:

$$L_{photometric}(I_{L_{warp}}, I_R) = \frac{1}{|M|} \sum_{m \in M} ||I_{L_{warp}}(m) - I_R(m)||_1,$$

where M is the mask that for all pixel index $m \in M$, $I_{L_{warp}}(m)$ is not zero.

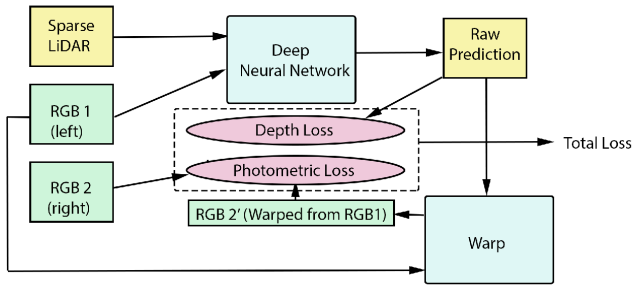

Figure 1: Training method.

Step 2: Disparity Refinement

Once a raw depth map is predicted in the first step, we will compute a disparity map using the equation: $z = \frac{fB}{d}$, where z and d represents the disparity and depth value, with f and B being the focal length and baseline of the stereo camera.

Next, given the left image I_L, the right image I_R, the raw disparity map D_{prior} and search range S, we will refine the disparity by:

$$D_{after}(i,j) = D_{prior}(i,j) + \arg\min_{s \in S} Cost(i,j,s),$$

where i and j are the row and column number of the point in the disparity map, and the cost function $Cost(i,j,s)$, which we currently use, is the simplest *sum of the absolute differences (SAD)*, which can be computed by:

$$Cost(i,j,s) =$$

$$\sum_{(i'j') \in N_{(i,j)}} |I_L(i',j') - I_R(i',j' + D_{prior}(i',j') + s)|,$$

, where $N_(i,j)$ represents a certain neighborhood of (i,j).

Thus, we get the final disparity map, D_{after}, as the final output of the framework.

Current Result & Future Work

Currently, we trained and tested our framework using the KITTI dataset (Geiger et al, 2012), and achieve a root mean squared error (RMSE) of 893, which is in the same level as the state-of-art (Wang et al, 2019).

In the future, we plan to improve our framework by adding more convolutional layers to our depth prediction network, implementing better cost functions to find the local optimum more accurately, and designing better refinement methods that take the disparities of neighbor points into consideration.

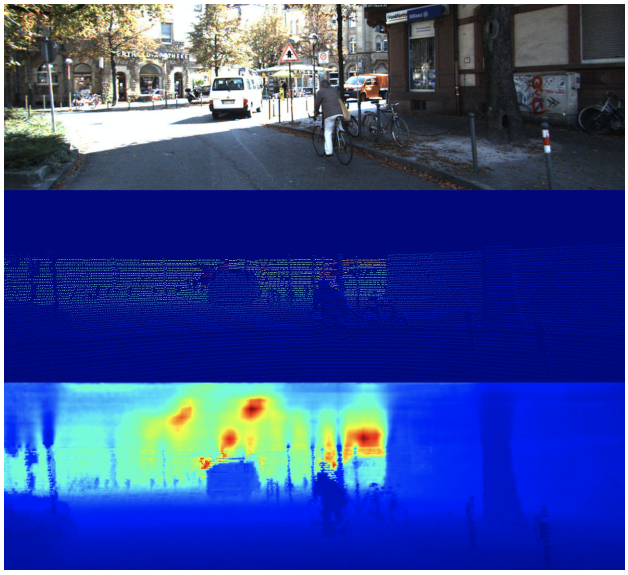

Figure 2: From top to down: Input color image (left), input sparse LiDAR depths, and depth map predicted by our method

Refrences

Wang et al (2019); Wang et al, 2019
Wang, T.; Hu, H.; Lin, C. H.; Tsai, Y.; Chiu, W.; and Sun, M. 2019. 3D LiDAR and Stereo Fusion using Stereo Matching Network with Conditional Cost Volume Normalization. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*

Kendall et al (2017)
Kendall, A.; Martirosyan, H.; Dasgupta, S.; Henry, P.; Kennedy, R.; Bachrach, A.; and Bry, A. 2017. End-to-End Learning of Geometry and Context for Deep Stereo Regression. In *Proceedings of The IEEE International Conference on Computer Vision (ICCV)*

Park et at (2018)
Park, K.; Kim, S.; and Sohn, K. 2018. High-precision Depth Estimation with the 3D LiDAR and Stereo Fusion. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*

Hirschmüller, 2008
Hirschmüller, H. 2008. Stereo Processing by Semi-Global Matching and Mutual Information. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*

Ma et al (2019)
Ma, F.; Cavalheiro, G. V.; and Karaman, S. 2019. Self-Supervised Sparse-to-Dense: Self-Supervised Depth Completion from LiDAR and Monocular Camera. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*

smallskip *Geiger et al ,2012*
Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2012. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*