Predicting LiDAR Intensity from RGB and Depth Images

Computer Science Project Report for Sensor Simulation Challenge

vorgelegt von

Carsten Schmotz

geb. am 23.09.1996 in Dachau

Department Informatik
Lehrstuhl Graphische Datenverarbeitung
Friedrich-Alexander-Universität Erlangen-Nürnberg

Betreuer: Richard Marcus

Betreuender Hochschullehrer: Prof. Dr. Marc Stamminger

Contents

1 Introduction **3**

1.1 Motivation . . . 3

1.2 Contribution . . . 3

1.3 Related work . . . 3

2 LiDAR Intensity Prediction **4**

2.1 Setup . . . 4

2.1.1 Bilateral Propagation Net (BP-Net) . . . 4

2.1.2 Depth-Anything Models . . . 4

2.1.3 Pix2pix Network . . . 4

2.2 Implementation . . . 5

2.2.1 Generation of depth maps . . . 5

2.2.2 Model Training . . . 6

2.3 Results . . . 7

2.3.1 Depth BP-Net to LiDAR . . . 7

2.3.2 Depth-Anything v1 to LiDAR . . . 8

2.3.3 Depth-Anything v2 to LiDAR . . . 9

2.3.4 Metric Depth to LiDAR . . . 10

2.3.5 RGB plus BP-Net to LiDAR . . . 11

2.3.6 RGB plus Depth-Anything v1 to LiDAR . . . 12

2.3.7 RGB plus Depth-Anything v2 to LiDAR . . . 13

2.3.8 RGB plus Metric Depth to LiDAR . . . 13

3 Conclusion **15**

4 Appendix **16**

4.0.1 RGB to LiDAR . . . 16

Abstract

This report explores the application of the Pix2Pix network for predicting LiDAR intensity maps using RGB images and depth information as additional input. The dense depth maps used are generated from the RGB images using the Bilateral Propagation Network for Depth Completion and the DepthAnything models (v1 and v2). This approach uses RGB images from the KITTI dataset and explores potential improvements in prediction accuracy and robustness over conventional methods.

Chapter 1

Introduction

1.1 Motivation

LiDAR sensors provide critical depth information for autonomous driving and robotics. LiDAR intensity maps are often sparse and incomplete. Using depth maps as additional input is a way to improve the richness and accuracy of LiDAR prediction. The Pix2Pix image-to-image translation network is a promising approach for integrating these modalities [7] [2].

1.2 Contribution

This project explores the use of the Pix2Pix network to predict LiDAR intensity maps using RGB images and depth maps as additional inputs.

1.3 Related work

BP-Net (Bilateral Propagation Network): BP-Net is a neural network architecture designed for depth completion tasks. It generates dense depth maps from sparse data [4].

Depth-Anything Models: Depth-Anything represents a significant advance in monocular depth estimation by using both labelled and unlabelled data on a large scale [6]. There are two versions available, the second having better accuracy and robustness of depth estimation [5]. These nets are also capable of performing metric depth to provide distance measurements from images.
3D Reconstruction Techniques: 3D reconstruction focuses on creating three-dimensional models from two-dimensional images or depth data. 3D reconstruction techniques include methods such as structure-from-motion (SfM) and multi-view stereo (MVS). These techniques aim to generate accurate 3D representations of scenes or objects from multiple views or depth sensors [3].

Chapter 2

LiDAR Intensity Prediction

2.1 Setup

First, the depth was created using RGB images from "The KITTI Dataset" [1]. In order to get a large variety of depths for comparison, the Bilateral Propagation Network (BP-Net), Depth-Anything v1 & 2 and Metric v2 were used. The metric depth v1 was not used because it did not run in time.

2.1.1 Bilateral Propagation Net (BP-Net)

- **Purpose:** Used for depth completion to improve depth maps from incomplete or noisy data.
- **Data set:** The input was raw data from kitti with a depth completion set already available.

2.1.2 Depth-Anything Models

- **Depth-Anything v1:** Provides initial depth estimates using large unlabeled data for zero-shot learning.
- **Depth-Anything v2:** Enhanced depth prediction capabilities, incorporating improvements over v1 for better depth map accuracy.
- **Metric Depth Estimation v2:** Enhanced depth maps with precise metric depth measurements.
- **Dataset:** Corresponding RGB images that were used in BP-Net for better comparison.

2.1.3 Pix2pix Network

- **Purpose:** Image-to-Image processing to predict LiDAR intensity from RGB and depth images.

- **Training Data:** RGB from the KITTI dataset and their created depth pairs. Eight different input data variations were used, see in section 2.3.

2.2 Implementation

2.2.1 Generation of depth maps

- **BP-Net:** An existing pre-trained model on Kitti data was loaded. The BP-Net model was applied to generate depth maps from the KITTI dataset. The output was 1000 depth maps.
- **Depthanthing v1 & 2 and metric depth:** The depth networks also provide trained models for kitti. There are three versions of all depth nets. The small, the base and the large model. For similar results the base model was loaded in every case.
- **Dataset preparation:** RGB images were paired with the generated depth maps to create a comprehensive training dataset for the pix2pix model.

The four depth maps sources (shown in figure 2.1) have different properties. The BP-Net has very little near-field resolution. The Depth-Anything depth maps are both more dense in the near field, but lack the distant objects. Version 2 has better edge visualisations and some distant objects like the sign on the left or the car are better visible. The metric depth has the best focus on the far field, but has very little information in near field.

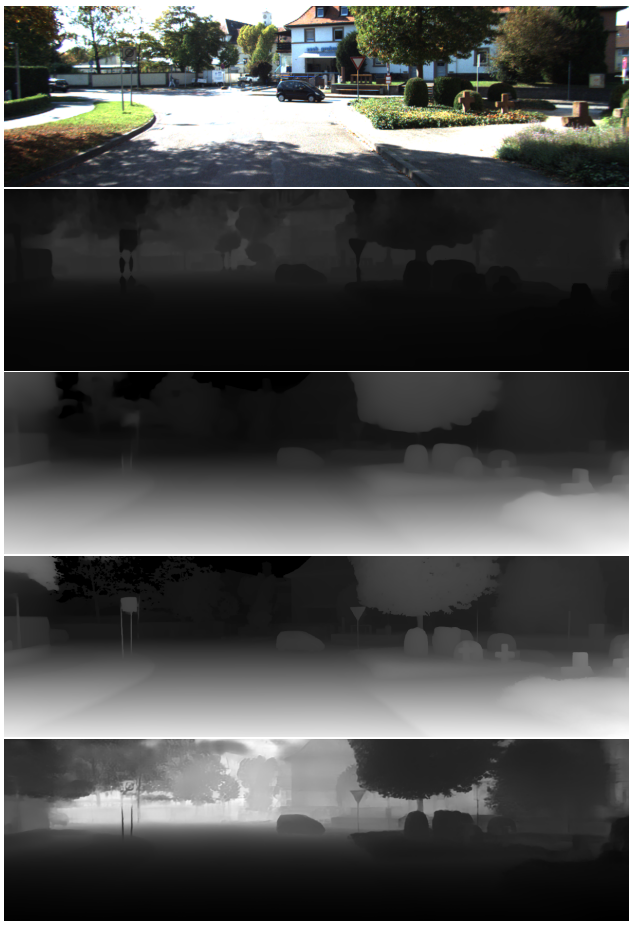


Figure 2.1: RGB and created depth (BP, Deptanything v1,2 and metric v2)

2.2.2 Model Training

- **Pix2pix configuration:** The pix2pix architecture was modified to accept RGB images and depth maps (4 dim.) as inputs to be able to scan the input channel seperately and to avoid losses due to overlaying RBG and depth images. The unmodified pix2pix was trained to predict LiDAR intensity values based on depth maps inputs. The 4 dimensional pix2pix was given the input of rgb plus depth maps.

- **Training process:** Training parameters such as learning rate 0.0002000, batch size 1, and number of epochs 20 were set. The model was trained using the prepared dataset in the configuration 1000 images divided into 800 training, 100 validation, 100 test folders. The input is different for the sole depth input and the RGBD input because the script for dividing are not executed in the same way. This was discovered after the training and test runs. Both versions were run with masked out the pixel value -1 to reduce the areas of no interest. The

input was cropped to a size of 256×256 but not scaled.

2.3 Results

This section presents the results of the experiments using different setups to predict LiDAR intensity. The performance of several methods was evaluated, including the separate depths from BP-Net, Depth-Anything v1 and v2, and metric depth estimation to LiDAR. The other four are RGB plus each of these depths to LiDAR. The images returned are crop and the full versions are shown for comparison. The cropped version is the right part of the full resolution (1216×352) images. For the 4 dimensional runs, the cropped LiDAR groundtruth is used for better visualisation as this pix2pix version gives these additional outputs. The results of eight runs are summarised below.

2.3.1 Depth BP-Net to LiDAR

Using only the depth maps generated by BP-Net (Figure 2.2), without accompanying RGB images, the model was trained to predict LiDAR intensity. This approach resulted in some visible predictions.

Figure 2.2: Depth BP-Net LiDAR and predicted LiDAR.

2.3.2 Depth-Anything v1 to LiDAR

The Depth-Anything v1 model (Figure 2.3) provided initial depth estimates using large-scale unlabelled data. When used to predict LiDAR intensity, this model showed results that are more in line with depth maps rather than LiDAR intensity.

Figure 2.3: Depth Depth-Anything v1 LiDAR and predicted LiDAR.

2.3.3 Depth-Anything v2 to LiDAR

The Depth-Anything v2 (Figure 2.4) shows almost the same result as v1. It could be that pix2pix was trying to learn LiDAR to depth, but the datasets and the training settings were correct for that.

Figure 2.4: Depth-Anything v2 LiDAR and predicted LiDAR.

2.3.4 Metric Depth to LiDAR

In this approach, accurate metric depth measurements (Figure 2.5) were used to predict LiDAR intensity. This method significantly shows a good prediction of LiDAR intensity.

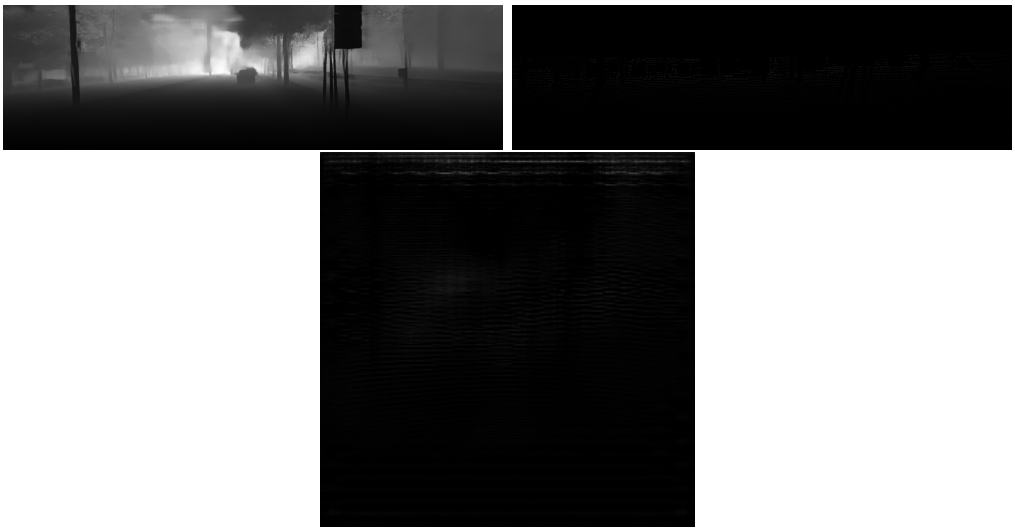


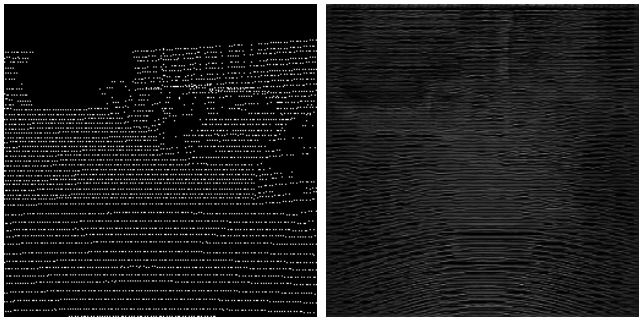
Figure 2.5: Metric depth LiDAR and predicted LiDAR.

2.3.5 RGB plus BP-Net to LiDAR

In this setup (Figure 2.6), RGB images were combined with depth maps generated by the BP-Net model to predict LiDAR intensity. The integration of depth information from BP-Net shows a significant increase in prediction accuracy compared to using depth map images alone. For example, the cars are barely visible and the background is almost incorrectly filled, perhaps due to masking.

(a) RGB Image

(b) Depth Map



(c) Groundtruth LiDAR **(d)** Predicted LiDAR

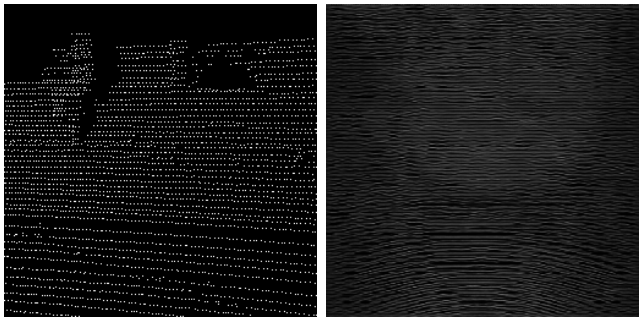
Figure 2.6: RGB, Depth, Predicted LiDAR, and Predicted LiDAR images for the BP-Net setup.

2.3.6 RGB plus Depth-Anything v1 to LiDAR

When RGB images were combined with depth maps from Depth-Anything v1 (Figure 2.7), the model performed better than using depth alone. The fusion of RGB and depth data provided a more comprehensive input, resulting in a LiDAR intensity that is close to the ground truth. The end of the road is visible.

(a) RGB Image

(b) Depth Map



(c) Groundtruth LiDAR **(d)** Predicted LiDAR

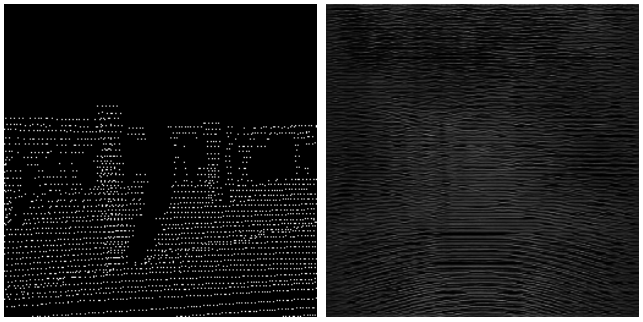
Figure 2.7: RGB, Depth, LiDAR, and Predicted LiDAR images for the Depth-Anything v1 setup.

2.3.7 RGB plus Depth-Anything v2 to LiDAR

RGB images paired with Depth-Anything v2 depth maps (Figure 2.4) show better prediction than v1. The sign in the ground truth is more visible due to the better depth map resolution of Depth-Anything v2.

(a) RGB Image

(b) Depth Map



(c) Groundtruth LiDAR **(d)** Predicted LiDAR

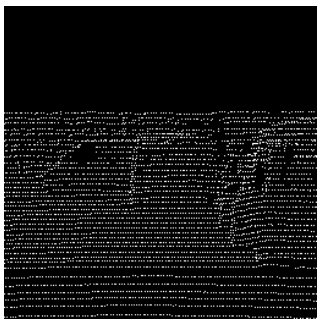
Figure 2.8: RGB, Depth, LiDAR, and Predicted LiDAR images for the Depth-Anything v2 setup.

2.3.8 RGB plus Metric Depth to LiDAR

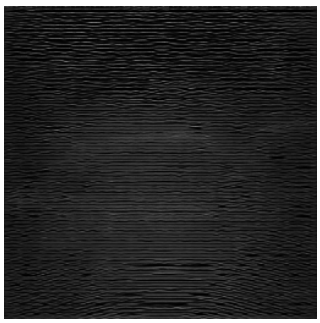
The combination of accurate metric (Figure 2.9) depth measurements produced the most accurate predictions of LiDAR intensity. This setup exploited the strengths of both RGB data and metric depth and gave the best results in all runs. The contours of the cars are visible on the right.

(a) RGB

(b) Depth



(c) Groundtruth LiDAR



(d) Predicted LiDAR

Figure 2.9: RGB, Depth-Anything metric, LiDAR and predicted LiDAR.

Chapter 3

Conclusion

In conclusion, the results of the different depth maps against LiDAR show poor prediction for most of the single depthmap inputs. The combined inputs perform better in prediction. The comparisons are not fully acceptable because there were problems with the different output images and mostly not matching samples. Also, the supplied script did not work for this evaluation because the cropped output image should be stitched together in pairs of 4 to get the full image. This makes it difficult to analyse the data.

Overall, there are some differences to note: Depth-Anything v2 performs better than v1, and metric depth is better than both. The BP network also shows some good results. A better approach is to use the same output images and run without masking for a more reliable result. Also, the lack of loss function due to non-working scripts and not tracked by wandb is a big disadvantage for evaluation.

Chapter 4

Appendix

4.0.1 RGB to LiDAR

These are the results for the pix2pix model trained using only RGB images to predict LiDAR intensity. The predicted LiDAR appears to be good, but to focus more on the different depths, this has been omitted.

Figure 4.1: RGB LiDAR and predicted LiDAR.

Bibliography

[1] Andreas Geiger et al. "Vision meets Robotics: The KITTI Dataset". In: *International Journal of Robotics Research (IJRR)* (2013).

[2] Phillip Isola et al. "Image-to-Image Translation with Conditional Adversarial Networks". In: *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. 2017.

[3] A. Saxena, S.H. Chung, and A.Y. Ng. "3-D Depth Reconstruction from a Single Still Image". In: *International Journal of Computer Vision* 76.1 (2008), pp. 53–69. DOI: 10.1007/s11263-007-0071-y.

[4] Jie Tang et al. "Bilateral Propagation Network for Depth Completion". In: *CVPR* (2024).

[5] Lihe Yang et al. "Depth Anything V2". In: *arXiv preprint arXiv:2406.09414* (2024).

[6] Lihe Yang et al. "Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024.

[7] Jun-Yan Zhu et al. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks". In: *Computer Vision (ICCV), 2017 IEEE International Conference on*. 2017.