Predicting LIDAR Intensity from RGB and Depth Images

Project report in computer science for sensor simultion 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

Beginn der Arbeit: 22.07.2024

Abgabe der Arbeit: 29.07.2024

Contents

Abstract

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

Chapter 1

Introduction

1.1 Motivation

LiDAR senors provide critical depth information for autonomous driving and robotics. The LiDAR intensity maps are often spares and incomplete. Using depht maps as an addtional input is a way to impove the richness and accuracy ot the LiDAR predection. The Pix2Pix network for image-to-image translation offers a promising approch for integrating these modalities.

1.2 Contribution

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

1.3 Related Work

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

DeptAnything Models: DepthAnything represents a significant advancement in monocular depth estimation by leveraging both labeled and unlabeled data at a large scale [6]. There are two version availble the second one has a better accuracy and robustness of depth estimaions [5]. Furthermore are these networks cabible to perform metric depth to provide distance measurements from images.

3D Reconstruction Techniques: 3D reconstruction focused on creating three-dimensional models from two-dimensional images or depth data. Techniques in 3D reconstruction include methods like 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

Predicting LiDAR Intensity

2.1 Setup

First of all the depth was created thru rgb pictures taken from "The KITTI Dataset" [1]. In order to get a great variatyof depth for comparison Bilateral Propagation Network (BP Net), Depthanything v1 & 2 and metric v2 was used. The metric depth v1 was not used because it did not run intime.

2.1.1 Bilateral Propagation Network (BP Net)

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

2.1.2 Depth Anything Models

- **Depth Anything v1:** Provided initial depth estimations using large-scale 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:** Supplemented depth maps with precise metric depth measurements.
- **Dataset:** Corresponding RGB images 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 [7] [2].

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

2.2 Implementation

2.2.1 Depth Map Generation

- **BP Net:** An existing pretrained model on kitti data was loaded. Utilized BP Net and Depth Anything models to generate depth maps from the KITTI dataset. Depth maps were processed to ensure consistency and accuracy before use. The output was 1000 depthmaps
- **Depthaynthing v1 & 2 and Metric Depth:** The depthaynthing networks provide also trained models for kitti. For all depthanything networks there aret three version availible. The smallm, base and large model. For similiar results the basemodel was loaded in each case.
- **Dataset Preparation:** RGB images were paired with the generated depth maps to create a comprehensive training dataset for the pix2pix model.

The three dephtmaps sources (shown in picture 2.1) have different properties. The BP Net has very little resultion in the nearfield. The depthanything depthmaps are both more dense in the nearfield but the far distant objects in backround are missing. The version 2 has better edge visualations and some distance objects like the sign on the lefthand side or the car are better visibel. The metric depth has the best focus on the farplane from all of them but near field has very little informtion.

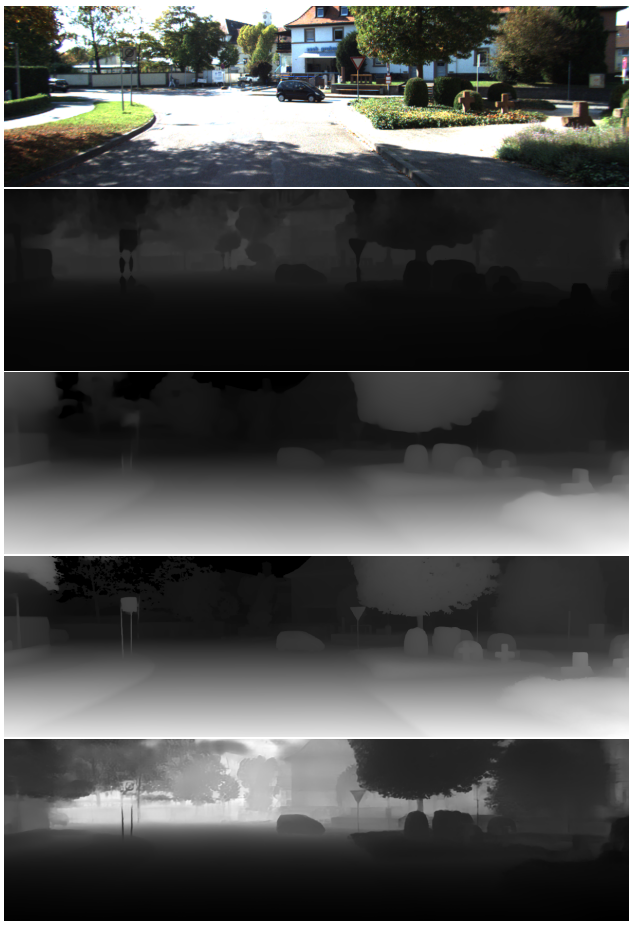


Figure 2.1: RGB and created depth (BP, Deptanything v1,2 and metric v2)

2.2.2 Model Training

- **pix2pix Configuration:** The pix2pix architecture was adapted to accept RGB images and depth maps (4 dim.) as inputs to be able to scan input channel seperately and to avoid loss by overlaying rgb and depth images. The unmodified pix2pix was trained to predict LiDAR intensity values based on depthmaps inputs. The 4 dimensional pix2pix was given the input of rgb plus depthmaps.

- **Training Process:** The 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 pictures divided in 800 train, 100 val, 100 test folders. The input is differnet for the sole depth input and the rgbd input due to not equal running scripts for dividing. This was discoverd after the training and testing runs. Both versions were executed with masking out the pixel value -1 to reduce the areas which are not of interest. The input

was crop to a size of 256 time 256 but not scale.

2.3 Results

This section presents the results of the experiments using different setups to predict LiDAR intensity. It was evaluated the performance of several methods, including the seperate depths from BP Net, Depth Anything v1 and v2, as well as metric depth estimation to LiDAR. The other four are RGB plus each of these depths to LiDAR. The returning pictures are crop and the full verions a used as comparison. The crop version is the right part of the full resolution (1216×352) images. For the 4 dimensional runs the cropped LiDAR groundtruth for better visualsation because this pix2pix version gives these extra outputs. The results from eight runs are summarized 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 yielded shows some visible prediction.

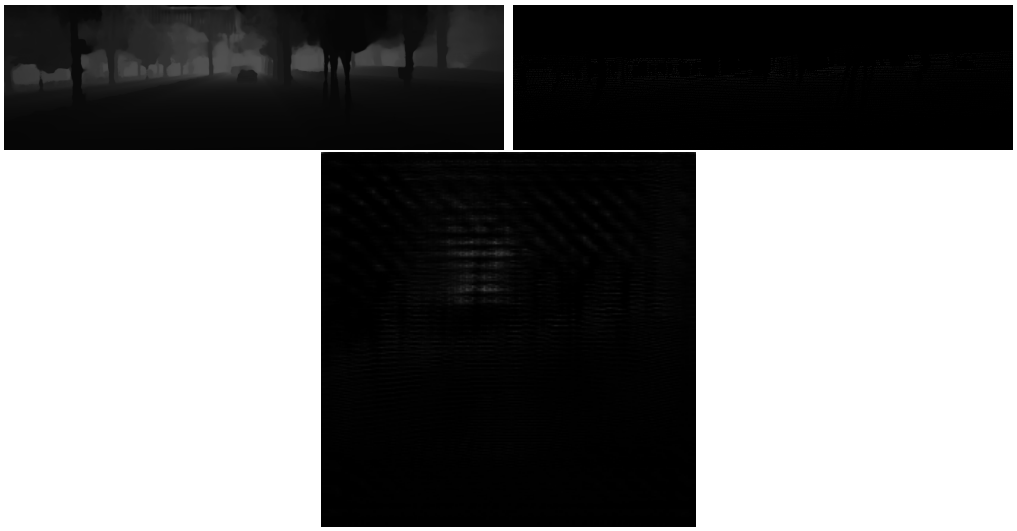


Figure 2.2: RGB LiDAR and predicted LiDAR.

2.3.2 Depth Anything v1 to LiDAR

The Depth Anything v1 model (Figure 2.3) provided initial depth estimations using large-scale unlabeled data. When used to predict LiDAR intensity, this model showed results that are more familiar with depth maps and not LiDAR intensity.

Figure 2.3: RGB LiDAR and predicted LiDAR.

2.3.3 Depth Anything v2 to LiDAR

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

Figure 2.4: Depthanything v2 LiDAR and predicted LiDAR.

2.3.4 Metric Depth to LiDAR

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

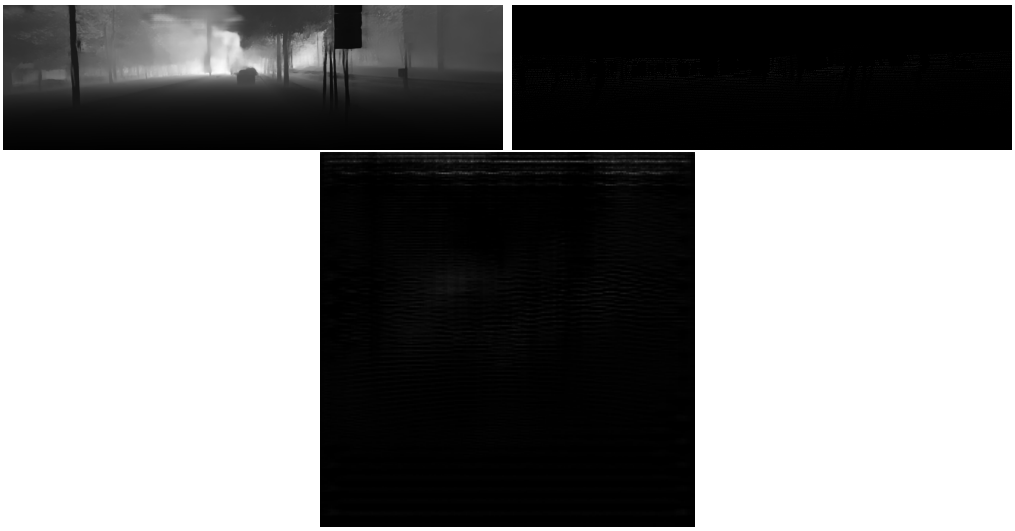


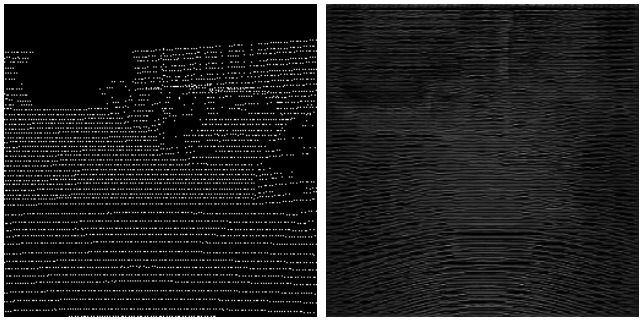
Figure 2.5: Metricdepth LiDAR and predicted LiDAR.

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 significantly displayes some correct prediction accuracy compared to using Depth maps images alone. The cars for instance are barely visibile and the backround is almost filled wrong perhaps due to the masking.

(a) RGB Image

(b) Depth Map



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

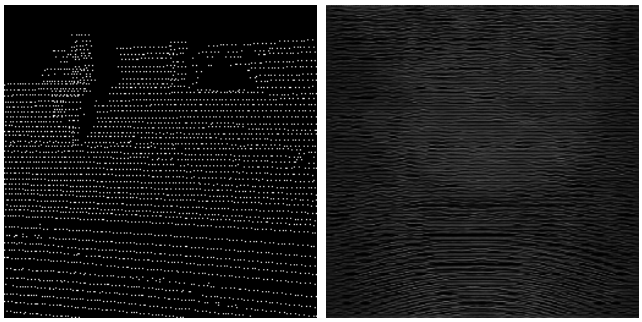
Figure 2.6: RGB, Depth, Predicted LiDAR, and Predicted LiDAR images for the BP Net setup.

2.3.6 RGB plus Depth Anything v1 to LiDAR

Combining RGB images with depth maps from Depth Anything v1 (Figure 2.7), the model achieved better performance than using depth alone. The fusion of RGB and depth data provided a more comprehensive input, resulting in an LiDAR intensity but it show reselemblest to the groundtruth. The end of the street is cleary visible.

(a) RGB Image

(b) Depth Map



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

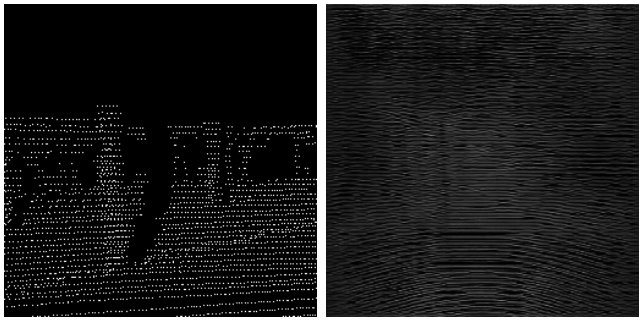
Figure 2.7: RGB, Depth, Predicted LiDAR, and Predicted LiDAR images for the Depth Anything v1 setup.

2.3.7 RGB plus Depth Anything v2 to LiDAR

RGB images paired with depth maps from Depth Anything v2 (Figure 2.4) show a better predection than v1. The sign in the groundtruth is more visible due to better depthmaps resolution of the depthanything v2.

(a) RGB Image

(b) Depth Map



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

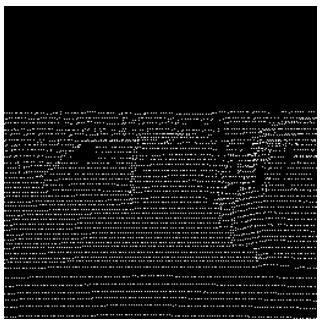
Figure 2.8: RGB, Depth, Predicted LiDAR, and Predicted LiDAR images for the Depth Anything v2 setup.

2.3.8 RGB plus Metric Depth to LiDAR

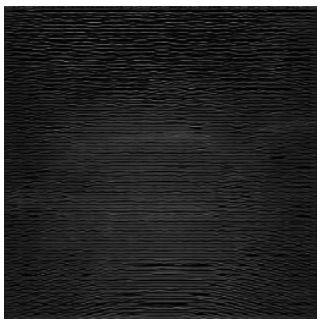
The combination of RGB images with precise metric (Figure 2.9)depth measurements led to the most accurate predictions of LiDAR intensity. This setup utilized the strengths of both RGB data and metric depth, achieving the best results across all runs. The conturs of the cars on the right side are visible.

(a) RGB

(b) Depth



(c) Groundtruth LiDAR



(d) Predicted LiDAR

Figure 2.9: RGB Depthanything metric LiDAR and predicted LiDAR.

Chapter 3

Conclusion

In conclusion, the results of different depthmaps against LiDAR are showing bad prediction for the for most of the sole depthmaps input. The combine inputs are performing better on the prediction. The comparions are not fully aceptalbe because there were problems with the different output images and mostly not machting samples. Furhtermore did the provide script not work for this installment because the cropped output picture should be in pairs of 4 be stited togehter to get the full image. This leads to an difficult evaltion on the data.

Overall there are some differnece to be notice dephtanything v2 performeces better dann v1 and the metric depth is better than both. The BP Net shows also some good result. A better approch is to use the same output images and to run without masking for a more reliable result. Furthermore is the missing of loss function due to not working scripts a big disadvantage for evalation.

3.1 Appendix

3.1.1 RGB to LiDAR

This are the results for the pix2pix model trained using only RGB images to predict LiDAR intensity. Predicted LiDAR seems to be fine but to concentrate more on the different depths this was left out.

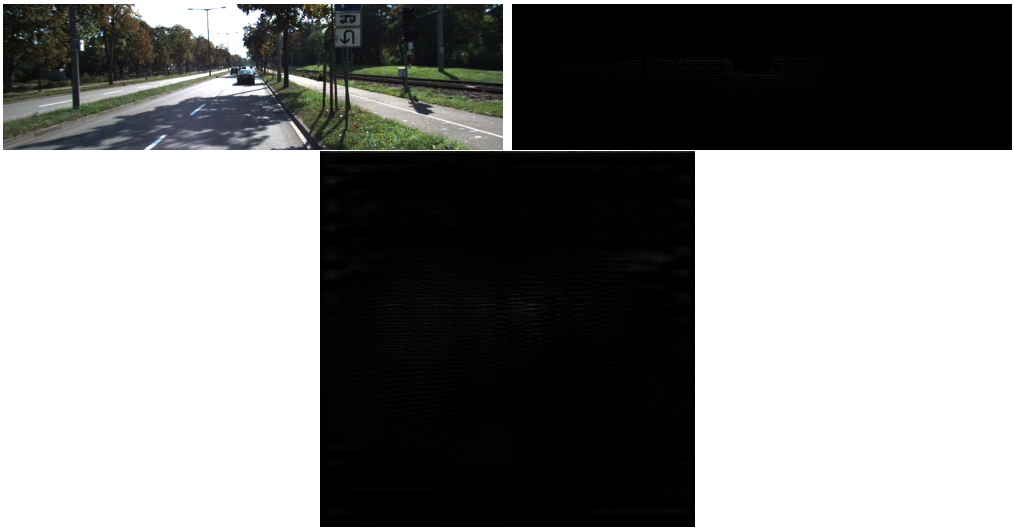


Figure 3.1: RGB LiDAR and predicted LiDAR.

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.