097200 - Deep Learning Course Final Project: Adaptive Depth Sampling

Gal Shitrit and Adam Wolff

May 2019

1 Introduction

In recent years, depth sensing has become essential for a variety of new significant applications. For example, depth sensors assist autonomous cars in navigation and in collision prevention [25]. The physical constraints on active depth sensing mobile devices, such as light detection and ranging (LiDAR), yield sparse depth measurements per scan. This results in a coarse point cloud and requires an additional estimation of missing data.

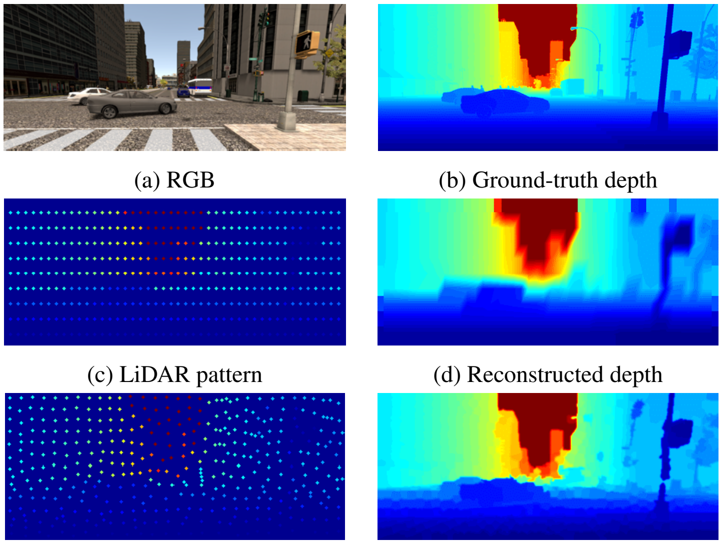
Traditional LiDARs have a restricted scanning mechanism. Those devices measure distance in specified angle intervals, using a fixed number of horizontal scan-lines (usually 16 to 64), depending on the number of transceivers. A new revolutionary technology is now emerging of solid-state depth sensors. They are based on optical phased-arrays with no mechanical parts, and can thus scan the scene fast in an adaptive manner (programmable scanning) [4, 21]. In addition, those innovative devices are much cheaper than those currently in use. This calls for the development of new, efficient, sampling strategies, which reduce the reconstruction error per sample. Since almost always autonomous platforms are equipped with RGB cameras, we investigate the possibility to improve the depth sampling process by taking the RGB information into account. Fig. 1 illustrates the task and clarifies our goal.

In this project, we address the topic of image-guided depth sampling. We use a deep neural network to find the optimal sampling locations for the task of depth reconstruction. Then, we demonstrate in experiments that our framework outperforms state-of-the-art depth completion methods for outdoor scenes.

2 Related Work

Depth completion: The task of depth reconstruction from scattered sparse samples is being increasingly investigated. The main methods can be divided to those which require only the sparse depth input (unguided) and to those assisted by additional information, e.g. color image (guided).

Among the unguided methods, some use classical approach [13, 17], while others rely on more advanced tools such as deep learning [6, 26].



(a) RGB

(b) Ground-truth depth

(c) LiDAR pattern

(d) Reconstructed depth

(e) Adaptive pattern

(f) Reconstructed depth

Figure 1: An illustration of the adaptive depth sampling task. Using the information gained from a color image (a), the sampling pattern can be modified to the scene (e) and achieve more accurate depth reconstruction (f) than simple LiDAR pattern (c-d)

On the contrary, guided methods exploit the connection between depth maps and their corresponding color image. Earlier methods used traditional image processing tools [3, 8]. Recently, several deep learning-based methods [5, 9, 11, 12, 14, 15, 18, 19] achieved state-of-the-art results.

Guided depth sampling: Despite the intensive development in depth completion, the issue of adaptive sampling is yet little addressed. Only [10, 16] have offered a non-trivial (i.e uniformly random or grid) sampling pattern as a previous step to depth reconstruction. Both studies selected sampling at locations which are most probable to have strong depth gradient. Nonetheless, they failed dealing with very low sampling budget of less than 5% of ground-truth pixels. Lately, A. Wolff presented in his M.Sc thesis a fast and practical image-guided algorithm for depth sampling and reconstruction, based on super-pixels. However, he didn't involve any learning technique (and particularly not a deep-learning approach) in his framework.

Nonuniform sampling: Over the years, the field of nonuniform sampling has been well established [1, 2, 20, 27]. However, these studies focus on the reconstruction of the signal for a given nonuniform sampling pattern and not on how to design data-driven patterns, given side information.

3 Method

3.1 Challenges

As explained before, given an RGB image of the environment we intend to find N points that will allow a minimal depth reconstruction error. The coordinates of the points can be represented in two ways:

1. as a set of N (x, y) pairs which correspond to the RGB image pixels.
2. as a binary image corresponding to the RGB image containing 1's in the pixels that should be sampled and 0's elsewhere.

Outputting a set is a difficult task. The main reason for this difficulty is related to the fact that sets are permutation invariant. Thus, well known regression loss functions such as $L1$ or mean squared error (MSE) will not work correctly. For instance, consider the following set of 1D points: $s_1 : \{1, 2, 3\}, s_2 : \{3, 2, 1\}$. Those sets are the same, but $L1$ loss will output 4 because it is not permutation invariant.

Outputting a binary image can lead to issues as well when using regression or cross-entropy loss functions. For instance, consider a binary image x_1 of N sampling points and an identical image x_2 in which all samples are shifted 1 pixel to the left. Those image will probably allow very similar depth reconstruction but $L1(x_1, x_2) = 2N$.

3.2 Architecture

We use a convolutional neural network (CNN) model in order to choose the desired N sampling points. The input of the model is an RGB image and the output is another image in which the pixel intensity represents the probability of a point to be sampled.

The model is based on the u-net segmentation network [23] (see Fig. 2). Originally, the input of the image is a grayscale image and the output is 2 channeled segmentation map. We changed the input layer of the network to input RGB image and the last layer of the network to output one channeled image with the same height and width as the input image.

We trained the network using binary cross-entropy (BCE) loss function. The loss function was calculated between the network output, which is a probability distribution map, and the ground truth sampling map, which can also be interpreted as a hard probability distribution map. For experimentation, we also trained the same network using MSE loss function. For this purpose, we transformed the ground truth hard distribution map into soft distribution map using a gaussian filter. The loss function was then calculated between the network output and the ground truth soft distribution map.

3.3 Post-processing

In order to convert this map to a sampling map, a non-max-suppression algorithm should be used. To do so, we used GMM clustering [22] on the probabilities image. The center of each of the Gaussians was chosen to be the sampling point, because it corresponds to the mean of each cluster. Note that ideally, to get maximal similarity

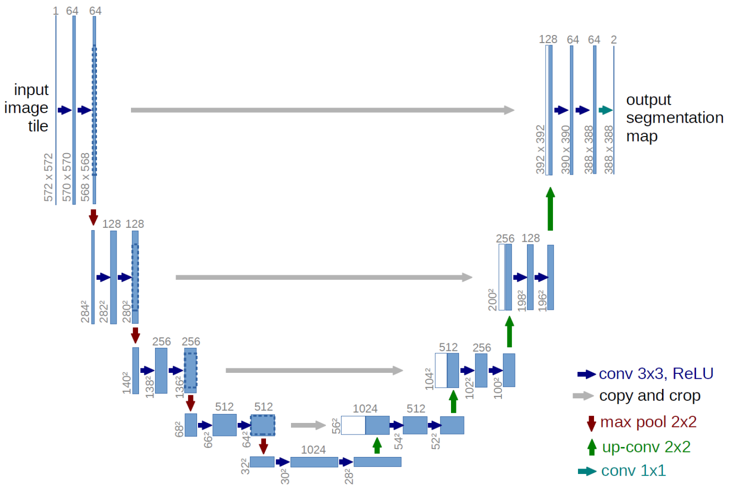


Figure 2: Original u-net architecture.

between the ground truth sampling map and the predicted sampling points, the loss should be computed after the fitting of the Gaussians. Unfortunately, this is not possible due to the fact that the above mentioned method is not differentiable - a property required for Gradient Descent optimization methods.

The predicted depth image is then reconstructed from the sparse depth image using interpolation. The interpolation can be performed with a linear or a higher order polynomial. A simpler reconstruction method is to assign for each pixel the depth value of the nearest neighbor from the sparse depth image. Note that the depth can be reconstructed more accurately using CNN [6] but this method is much more computationally expensive. A high-level block-diagram of our method is illustrated in Fig. 3.

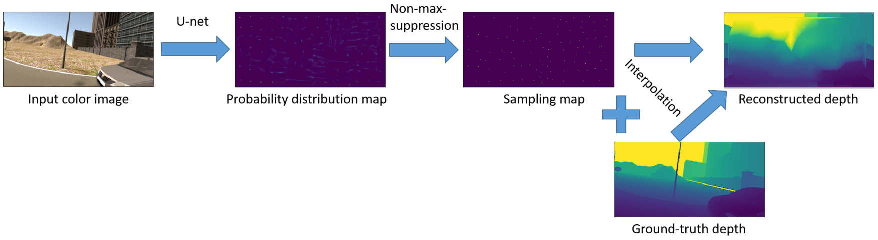


Figure 3: High-level block-diagram of our method.

4 Experiments

4.1 Dataset

The Synthia dataset [24] provides synthetic RGB, depth and semantic images for urban driving scenarios. We use synthetic data as for now there is no large non-synthetic dataset that provides a dense and accurate depth map. We need a dense depth map to be able to sample at any given point. Accuracy is required especially to show the increased resolution we obtain. Thus, two large real-life datasets do not apply: KITTI depth completion benchmark [26] has semi-dense depth, and Cityscapes [7] has low resolution depth, which is in some cases inaccurate. We randomly split the 6,296 images in summer sequence 5 to 6,000 for training and 296 for evaluation. All color and depth images are downsampled from originally 1280×640 pixels to 320×160 pixels.

Additionally, we implement the algorithm described in the M.Sc thesis of A. Wolff for 100 samples and use the received sparse sampling maps as ground-truth for training and testing.

4.2 Evaluation

To evaluate the benefit of our chosen sampling location, we apply a depth reconstruction over the resulting samples. Then, we measure the root mean squared error (RMSE) between the reconstructed and the ground-truth depth. We chose to make the reconstruction of the sparse depth samples by applying linear interpolation (with nearest-neighbor extrapolation) or nearest-neighbor (NN) interpolation, since other sophisticated methods are much more computationally expensive. We make the evaluation on 0-100m depth range, which is similar to the range of a typical vehicle-mounted LiDAR.

4.3 Results

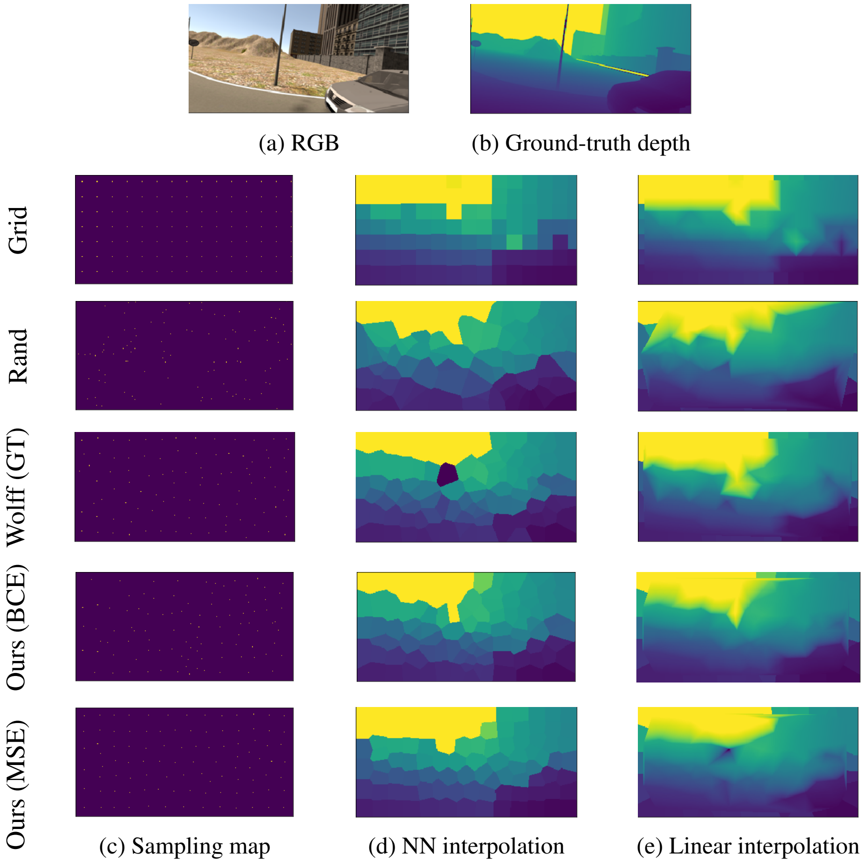
Quantitative results are presented in Table 1. Using our method, the resulting reconstructed depth is the second most accurate compared to other sampling methods, after Wolff's methods which we used as ground truth. Among our two proposed approaches, the one trained over BCE loss achieves better performance. Qualitative results, including sampling maps and NN and linear interpolated depth, are presented in Fig. 4.

5 Conclusion

In this project, we introduced a novel approach for image-based sparse depth sampling. We presented state-of-the-art results compared to traditional and modern sampling methods. We believe that this new direction calls for additional extensive research, in order to develop advanced, cheap and accurate depth sensing systems. Future work could Exploit temporal redundancy or use semantic information to further improve the performance and accuracy.

Sampling	Interpolation	RMSE [m]
Grid	NN	56.09
Random	NN	65.37
Wolff (GT)	NN	**44.48**
Ours (BCE)	NN	47.24
Ours (MSE)	NN	50.70
Grid	Linear	45.25
Random	Linear	55.84
Wolff (GT)	Linear	**43.01**
Ours (BCE)	Linear	43.41
Ours (MSE)	Linear	43.52

Table 1: Quantitative comparison for different depth sampling and reconstruction methods. Our method achieves second best result after ground-truth method.



(a) RGB (b) Ground-truth depth

(c) Sampling map (d) NN interpolation (e) Linear interpolation

Figure 4: Qualitative results: an example of a RGB image (a) and its ground truth depth image (b), following sampling map (c) and interpolated depths using nearest interpolation (d) and linear interpolation (e) of different sampling methods.

References

[1] Akram Aldroubi and Karlheinz Gröchenig. Nonuniform sampling and reconstruction in shift-invariant spaces. *SIAM review*, 43(4):585–620, 2001.

[2] Prabhu Babu and Petre Stoica. Spectral analysis of nonuniformly sampled data–a review. *Digital Signal Processing*, 20(2):359–378, 2010.

[3] Jonathan T Barron and Ben Poole. The fast bilateral solver. In *European Conference on Computer Vision*, pages 617–632. Springer, 2016.

[4] Pavel Cheben, Robert Halir, Jens H Schmid, Harry A Atwater, and David R Smith. Subwavelength integrated photonics. *Nature*, 560(7720):565, 2018.

[5] Zhao Chen, Vijay Badrinarayanan, Gilad Drozdov, and Andrew Rabinovich. Estimating depth from rgb and sparse sensing. *arXiv preprint arXiv:1804.02771*, 2018.

[6] Nathaniel Chodosh, Chaoyang Wang, and Simon Lucey. Deep convolutional compressed sensing for lidar depth completion. *arXiv preprint arXiv:1803.08949*, 2018.

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[8] Gilad Drozdov, Yevgengy Shapiro, and Guy Gilboa. Robust recovery of heavily degraded depth measurements. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 56–65. IEEE, 2016.

[9] Abdelrahman Eldesokey, Michael Felsberg, and Fahad Shahbaz Khan. Propagating confidences through cnns for sparse data regression. *arXiv preprint arXiv:1805.11913*, 2018.

[10] Simon Hawe, Martin Kleinsteuber, and Klaus Diepold. Dense disparity maps from sparse disparity measurements. In *13th International Conference on Computer Vision*, 2011.

[11] Zixuan Huang, Junming Fan, Shuai Yi, Xiaogang Wang, and Hongsheng Li. Hms-net: Hierarchical multi-scale sparsity-invariant network for sparse depth completion. *arXiv preprint arXiv:1808.08685*, 2018.

[12] Maximilian Jaritz, Raoul De Charette, Emilie Wirbel, Xavier Perrotton, and Fawzi Nashashibi. Sparse and dense data with cnns: Depth completion and semantic segmentation. In *2018 International Conference on 3D Vision (3DV)*, pages 52–60. IEEE, 2018.

[13] Jason Ku, Ali Harakeh, and Steven L Waslander. In defense of classical image processing: Fast depth completion on the cpu. *arXiv preprint arXiv:1802.00036*, 2018.

[14] Yaoxin Li, Keyuan Qian, Tao Huang, and Jingkun Zhou. Depth estimation from monocular image and coarse depth points based on conditional gan. In *MATEC Web of Conferences*, volume 175, page 03055. EDP Sciences, 2018.

[15] Yiyi Liao, Lichao Huang, Yue Wang, Sarath Kodagoda, Yinan Yu, and Yong Liu. Parse geometry from a line: Monocular depth estimation with partial laser observation. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5059–5066. IEEE, 2017.

[16] Lee-Kang Liu, Stanley H Chan, and Truong Q Nguyen. Depth reconstruction from sparse samples: Representation, algorithm, and sampling. *IEEE Transactions on Image Processing*, 24(6):1983–1996, 2015.

[17] Fangchang Ma, Luca Carlone, Ulas Ayaz, and Sertac Karaman. Sparse depth sensing for resource-constrained robots. *arXiv preprint arXiv:1703.01398*, 2017.

[18] Fangchang Ma, Guilherme Venturelli Cavalheiro, and Sertac Karaman. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. *arXiv preprint arXiv:1807.00275*, 2018.

[19] Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8. IEEE, 2018.

[20] Farokh Marvasti. *Nonuniform sampling: theory and practice*. Springer Science & Business Media, 2012.

[21] Christopher V Poulton, Ami Yaacobi, David B Cole, Matthew J Byrd, Manan Raval, Diedrik Vermeulen, and Michael R Watts. Coherent solid-state lidar with silicon photonic optical phased arrays. *Optics letters*, 42(20):4091–4094, 2017.

[22] Carl Edward Rasmussen. The infinite gaussian mixture model. In *Advances in neural information processing systems*, pages 554–560, 2000.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[24] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.

[25] Brent Schwarz. Lidar: Mapping the world in 3d. *Nature Photonics*, 4(7):429, 2010.

[26] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. *arXiv preprint arXiv:1708.06500*, 2017.

[27] J Yen. On nonuniform sampling of bandwidth-limited signals. *IRE Transactions on circuit theory*, 3(4):251–257, 1956.