Tackling Disturbed Depth Maps by Learning Input Data Confidence*

Abdelrahman Eldesokey
Linköping University
Linköping, Sweden
abdel62@liu.se

Mikael Persson
Linköping University
Linköping, Sweden
mikael.persson@liu.se

Michael Felsberg
Linköping University
Linköping, Sweden
michael.felsberg@liu.se

Fahad Shahbaz Khan
Inception Institute of Artificial Intelligence
Abu Dhabi, UAE
fahad.khan@inceptioniai.org

Abstract

The prevalence of affordable depth sensors with sparse output has led to numerous applications in autonomous driving and robotics. However, the data acquisition process of these sensors is prone to errors and disturbances which produces disturbed sparse depth maps. Deep learning based approaches have been investigated to produce rectified dense depth maps. Existing methods either utilize auxiliary information requiring huge networks or they fully trust the input with no correction mechanism. In this paper, and different to the existing approaches, we show that learning input confidence maps directly from the input and propagating it through the network is computationally more efficient and leads to significantly better results than existing methods. In addition, our approach is unsupervised, does not require any auxiliary information and utilizes a very compact network that makes it suitable for embedded applications. Experiments on the KITTI-Depth dataset show that our method achieves state-of-the-art results on the unguided scene depth completion. Our approach acts similar to nonlinear filtering and produces sharp edges and smooth surfaces. Finally, we show that our approach generalizes to other tasks by applying it to the problem of outlier rejection in sparse optical flow.

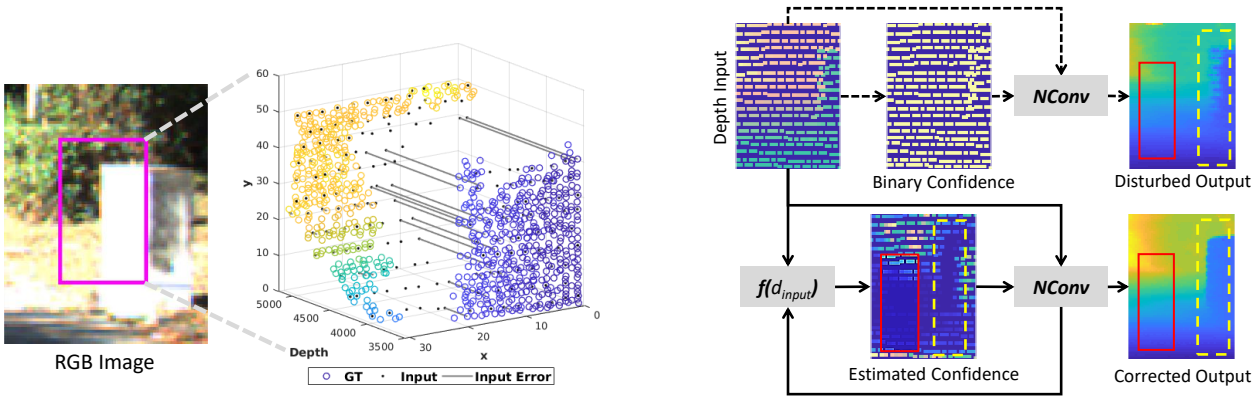
1. Introduction

In recent years, vision sensors capable of measuring depth information have gained popularity, leading to numerous advances in autonomous driving and robotics. Depth information enhances the understanding of the environment and facilitates decision support. Existing depth sensors, such as time-of-flight sensors and LiDARs, are either expensive or noticeably noisy and sparse. In addition, certain sensors or data acquisition mechanisms encompass complicated pipelines, which makes it difficult to validate the captured data. This is evident in the KITTI-Depth dataset [15] that contains a systematic error in the data due to the displacement between the RGB camera and the LiDAR [13] as illustrated in Figure 1a.

Thanks to advances in deep learning, several works have been proposed to address the problem of *scene depth completion*. In this problem, the sparse and noisy depth input produced by the depth sensors is rectified and mapped to a reliable dense output. Typically, deep learning methods employ different schemes to fuse different modalities, *e.g.* depth, RGB images [7, 11, 8, 4, 1, 16], and surface normals [13]. However, this is achieved using huge networks at the cost of enormous requirements regarding computations and memory, which are infeasible in autonomous driving and robotics applications. In contrast, sparsity-aware methods [5, 15] are very compact and they utilize only the depth modality. However, these methods fully trust the input and have no correction mechanism for the disturbed input as shown in Figure 1b.

In this paper, we propose an efficient approach for tackling the problem of disturbed depth maps by learning the input confidence directly from the disturbed input and propagating it through the network. Our proposed approach is integrated with a normalized convolutional neural network (NC) [5, 4] and trains end-to-end in unsupervised manner to perform unguided scene depth completion. In addition, our approach does not require any auxiliary information or groundtruth and estimates the input confidence based solely on the network error. Experiments shows that our approach sets a new-state-of-the-art on the unguided depth completion task of the KITTI-Depth dataset [15] and significantly

*The first workshop on Statistical Deep Learning for Computer Vision, in *Seoul, Korea*, 2019.



(a) (b)

Figure 1: **(a)** The KITTI-Depth dataset [15] contains systematic errors. The input (dots) and the groundtruth (circles) do not agree at occluded regions. **(b)** Existing confidence-aware deep networks, *e.g.* the normalized convolutional neural network (NC) [5], assume full correctness of the sparse input. This causes the reconstructed output to become erroneous (top yellow box). Our approach efficiently learns the input confidence in an unsupervised manner directly from the depth input, without requiring any auxiliary information. As a result, the disturbed measurements are accurately filtered out (bottom yellow box), leading to an improved depth completion.

improves over the standalone NC (see Figure 1b). Moreover, our approach significantly improves the results by learning the extent to which each sample in the input should contribute to the output similar to nonlinear filtering. This is demonstrated by evaluating our approach on the synthetic error-free dataset virtual-KITTI [6], Finally, we show the applicability of our approach to other problems by performing an experiment on a sparse optical flow task. The experiment shows that our approach successfully identifies noisy flow fields and is able to rectify them.

2. Related Work

The problem of scene depth completion received increased attention recently thanks to the availability of benchmarks, such as the KITTI-Depth [15] dataset. Several deep learning-based approaches were introduced in recent years that tackle the scene depth completion problem by exploiting the information in the available modalities. The majority of methods utilize an encoder-decoder architecture to deal with sparsity and the disturbed input concurrently. Ma *et al*. [11] proposed a deep regression network, which employs an encode-decoder architecture to fuse depth and RGB modalities. Similarly, Jaritz *et al*. [8] evaluated different fusion schemes for an encoder-decoder architecture. Another approach is to process each modality individually and learn confidence/uncertainty measures to fuse them. Qui *et al*. [13] learns an occlusion mask from the RGB image and utilize this mask to fuse the disturbed depth with surface normals. Gansbeke *et al*. [16] processes the RGB and the depth modality in two streams and learns confidence maps that are used to fuse the streams.

The aforementioned approaches utilize huge networks with high computational cost. An alternative compact approach for scene depth completion is the sparsity-aware deep networks, which accompany the sparse input with a confidence mask to describe the reliability of each sample point in the input. Uhrig *et al*. [15] introduced sparsity invariant convolution, which utilizes validity masks to identify valid pixels in the input. Eldesokey *et al*. [5] introduced the normalized convolutional neural network (NC), which replaces validity masks with continuous confidence masks that are propagated between layers. They also proposed different fusion schemes with the RGB images in [4]. However, these approaches assume full or known confidences at locations in the input where data exist. This assumption is infeasible since real data can be corrupted with disturbed measurements due to noise or calibration errors.

Our Approach: Different to the aforementioned approaches, we learn a continuous-valued input confidence in an unsupervised manner directly from the sparse input. Our proposed approach is *fully unguided* and does not require any type of auxiliary data, *e.g.* RGB images, as in [13, 16]. Further, our approach trains in an end-to-end fashion jointly with the NC network [5]. Our proposed approach is computationally efficient *and* achieves state-of-the-art results on the unguided depth completion task on the KITTI-Depth dataset [15]. We further demonstrate that our approach improves the reconstruction even on flawless synthetic datasets and that it generalizes to other domains by applying it to noisy sparse optical flow data.

3. Method

A typical deep learning approach for scene depth completion tries to learn a dense output $\mathbf{y} = f(\mathbf{x})$, where $\mathbf{x}$ is the sparse input depth in our case, and f is the deep network. Confidence-aware approaches [5, 4, 13, 16] on the other hand tries to learn a dense output aside with an output confidence map $(\mathbf{y}, \mathbf{c}_o) = f(\mathbf{x}, \mathbf{c}_i)$, where $\mathbf{c}_i$ is the binary input confidence map and $\mathbf{c}_o$ is the estimated output confidence map. In our proposed approach, we learn the input confidence map that minimizes the error at the dense output, $(\mathbf{y}, \mathbf{c_o}) = f(\mathbf{x}, h(\mathbf{x}))$, where h is a task-dependent input confidence estimation network.

To be able to learn the input confidence estimator h that minimizes the error, the network needs be able to propagate confidence between layers. The normalized convolutional neural network (NC) [5] satisfies this property as it propagates the input confidence between layers to produce an output confidence in conjunction with the dense output. Here, we replace the binary input confidence from the sparse input with the confidence estimated using our confidence estimation network.

3.1. Proposed Confidence Estimator

In our attempt to design the input confidence estimation network, we choose a compact U-Net architecture [14] to learn the mapping between the sparse depth input and the continuous-valued input confidence. The U-Net has been extensively used to exploit multi-scale information with great success [14]. This property will facilitate the learning of the confidence by exploiting each neighborhood at different scales and efficiently fusing them. Since the computational demands are an important factor in our work, we only perform three downsampling/upsampling steps instead of four. In addition, we use a small number of channels per layer to lower the memory requirements.

One crucial aspect when we design our confidence estimator is that confidences are non-negative. This constrains the choice of the activation function at the last layer to produce only non-negative values. Therefore, we use the sigmoid function as an activation at the last layer of the confidence estimator. Our network is trained using the mean L1 norm of the difference between the prediction and the groundtruth where available.

4. Experiments

To demonstrate the capabilities of our proposed method, we evaluate it on the KITTI-Depth benchmark [15]. Further, we evaluate on the synthetic virtual-KITTI dataset [6] to validate that our proposed method has an advantage even when the data is perfect. Finally, to demonstrate the generalization capabilities of our approach to other tasks, we apply it to the task of outlier rejection in sparse optical flow.

	MAE [mm]	RMSE [mm]	iMAE [1/km]	iRMSE [1/km]
SparseConv [15]	481.27	1601.33	1.78	4.94
ADNN [2]	439.48	1325.37	3.19	59.39
NN+CNN [15]	416.14	1419.75	1.29	3.25
NC [5]	360.28	1268.22	1.52	4.67
S2D [11]	288.64	954.36	1.35	3.21
HMS-Net (d) [7]	258.48	*937.48*	1.14	2.93
Gansbeke *et al*. [16]	249.11	**922.93**	1.07	2.80
Spade [8]	248.32	1035.29	**0.98**	**2.60**
NC-Conf (Ours)	**228.53**	988.57	*1.00*	*2.71*
NC-Conf-C (Ours)	*232.75*	1034.86	1.02	4.58

Table 1: Quantitative results for *unguided* depth completion on the *test* set of the KITTI-Depth dataset [15]. **Bold** is the best, *Italic* is the second best.

4.1. Experimental Setup

The ADAM optimizer was used with a learning rate of 10^{-3}, and a batch size of 4. The PyTorch source code for our proposed approach will be made publicly available on Github. Different methods are evaluated based on the Mean Average Error (MAE), Root Mean Square Error (RMSE), iMAE, and iRMSE computed for the disparity.

4.2. Results for the KITTI-Depth Dataset

Table 1 shows quantitative results of the *unguided* methods in comparison on the test set of the KITTI-Depth dataset. Our proposed approach NC-Conf outperforms all other methods on three out of four metrics when compared individually except for Spade [8], where NC-Conf is better on MAE and RMSE. The results also show that our proposed confidence estimation approach provides a performance boost of about 45% compared to the standalone NC [5]. Moreover, the compact version *NC-Conf-C* with only 84k parameters still outperforms all other methods with respect to MAE. Note that a lower MAE indicates a better reconstruction of nearby objects which practically are important in autonomous driving.

We provide a qualitative example from the KITTI-Depth dataset in Figure 2, which shows the improvement over the NC [5]. The figure clearly shows severe artifacts produced by [5] (last row), especially along edges, due to use of wrong confidence information. On the other hand, our proposed method successfully mitigates these artifacts by learning the correct input confidence, which leads to very crisp edges. In addition, our proposed method produces smoother surfaces as it down-weights intermediate samples lying on a plane (shown in dark blue in the 3rd row in Figure 2) and uses end-points to fit the entire plane.

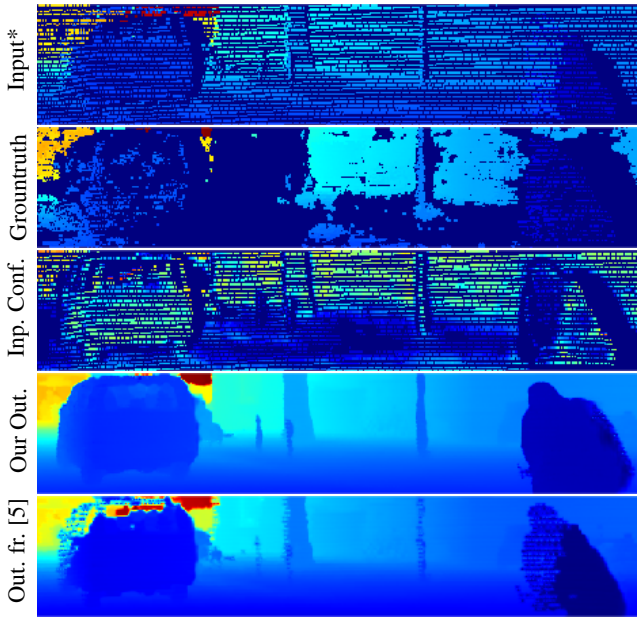


Figure 2: A qualitative example from the KITTI-Depth [15] dataset.

	RMSE	MAE	MRE	δ_1	δ_2	δ_3
NC [5]	8.27	1.87	0.0482	0.631	0.772	0.826
NC-Conf	7.26	1.53	0.0376	0.704	0.812	0.853

Table 2: Quantitative results for the virtual-KITTI dataset [6]. Note that δ is the inliers ratio as described in [3].

4.3. The Virtual KITTI Dataset

Synthetic datasets lack noise and therefore serve to show how the system performs when the input is perfect. As mentioned earlier, the confidence estimator also allows the system to perform adaptive smoothing similar to nonlinear filtering. This is useful because, even for noise free sparse data, not all values equally well predict their surroundings. Table 2 shows the clear improvement over the standalone NC (a performance boost of about 20%). The decrease of performance boost from 45% to 20% can be explained by the absence of disturbances/noise in the virtual-KITTI.

4.4. Generalization to Other Problems

We show that our approach generalizes to other problems by applying it to outlier rejection and interpolation of sparse optical flow dataset of our own making. This dataset contains 5k sequential stereo images and has flow vectors generated by KLT. The groundtruth is produced by geometrical verification over several frames, using a system based on [12].The optical flow components δ_x, δ_y are independently processed by our network and then concatenated. The network is trained using the L1 norm of the displacement vector error as loss. Quantitative results show an significant improvement from a MAE of **0.77** pixels when using the NC to **0.25** pixels when using our proposed approach. Figure 3 shows a qualitative result demonstrating the outlier removal and interpolation our approach.

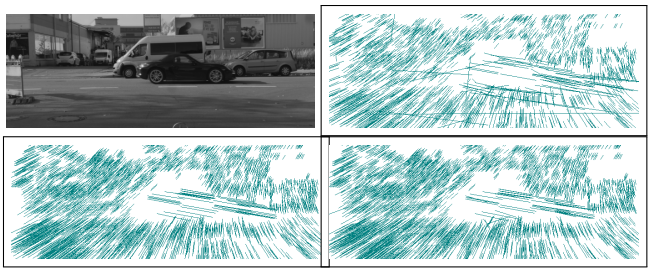


Figure 3: Qualitative example for optical flow outliers rejection. In right-bottom order, RGB frame, raw flow input, groundtruth flow, and estimated flow. Our method removes outliers and densifies the flow.

4.5. Discussion

We have shown that confidence is a useful measure of how sparse data should be weighted for specific task. Our estimated confidence successfully rejects disturbed samples in the KITTI-Depth and the optical flow experiment. Moreover, it weighs all samples depending how far they should be interpolated, producing smooth surfaces and crisp edges in all experiments.

The reason why we separate confidence estimation and interpolation is that interpolating is a sparse problem, whereas estimating confidence maps for sparse data is a local and dense problem. The absence of data is informative, and even for a network with a small receptive field, the confidence can be reliably estimated. Without separation, *e.g.* in depth completion, huge networks are required to simultaneously handle sparsity and noisy input.

5. Conclusion

We proposed an unsupervised approach to tackle the problem of disturbed depth maps by learning an input confidence map that is propagated through the network. Our approach, when integrated with the NC framework, is trained end-to-end and learns to identify disturbed measurements. We showed that our approach sets a new state-of-the-art on unguided depth completion on the KITTI-Depth benchmark at very low computational cost. We further show that the method improves over the standalone NC for undisturbed data using a sparsified virtual-KITTI dataset. Finally, we showed that our approach can be generalized by evaluating it on a real sparse optical flow dataset.

References

[1] X. Cheng, P. Wang, and R. Yang. Depth estimation via affinity learned with convolutional spatial propagation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–119, 2018.

[2] N. Chodosh, C. Wang, and S. Lucey. Deep Convolutional Compressed Sensing for LiDAR Depth Completion. mar 2018.

[3] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[4] A. Eldesokey, M. Felsberg, and F. S. Khan. Confidence propagation through cnns for guided sparse depth regression. *arXiv preprint arXiv:1811.01791*, 2018.

[5] A. Eldesokey, M. Felsberg, and F. S. Khan. Propagating confidences through cnns for sparse data regression. In *The British Machine Vision Conference (BMVC), Northumbria University, Newcastle upon Tyne, England, UK, 3-6 September, 2018*, 2018.

[6] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016.

[7] Z. Huang, J. Fan, S. Yi, X. Wang, and H. Li. HMS-Net: Hierarchical Multi-scale Sparsity-invariant Network for Sparse Depth Completion. *ArXiv e-prints*, Aug. 2018.

[8] M. Jaritz, R. de Charette, E. Wirbel, X. Perrotton, and F. Nashashibi. Sparse and dense data with cnns: Depth completion and semantic segmentation. *arXiv preprint arXiv:1808.00769*, 2018.

[9] J. Ku, A. Harakeh, and S. L. Waslander. In defense of classical image processing: Fast depth completion on the cpu. *arXiv preprint arXiv:1802.00036*, 2018.

[10] F. Ma and S. Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. *arXiv preprint arXiv:1709.07492*, 2017.

[11] F. Ma, G. Venturelli Cavalheiro, and S. Karaman. Self-supervised Sparse-to-Dense: Self-supervised Depth Completion from LiDAR and Monocular Camera. *ArXiv e-prints*, July 2018.

[12] M. Persson, T. Piccini, M. Felsberg, and R. Mester. Robust stereo visual odometry from monocular techniques. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, pages 686–691. IEEE, 2015.

[13] J. Qiu, Z. Cui, Y. Zhang, X. Zhang, S. Liu, B. Zeng, and M. Pollefeys. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. *arXiv preprint arXiv:1812.00488*, 2018.

[14] O. Ronneberger, P. Fischer, and T. 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.

[15] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger. Sparsity invariant cnns. In *International Conference on 3D Vision (3DV)*, 2017.

[16] W. Van Gansbeke, D. Neven, B. De Brabandere, and L. Van Gool. Sparse and noisy lidar completion with rgb guidance and uncertainty. *arXiv preprint arXiv:1902.05356*, 2019.