CNN-based synthesis of realistic high-resolution LiDAR data

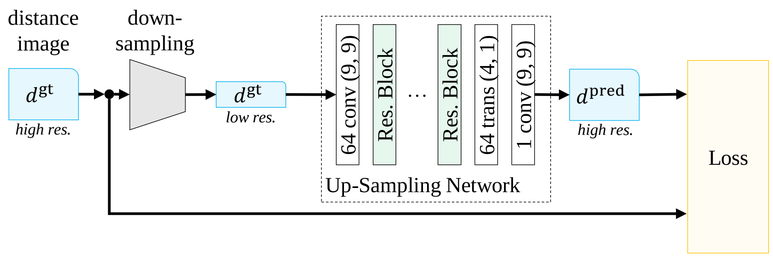
Larissa T. Triess[1,2], David Peter[1], Christoph B. Rist[1], Markus Enzweiler[1] and J. Marius Zöllner[2,3]

[1]Daimler AG [2]Karlsruhe Institute of Technology [3]FZI Research Center for Information Technology

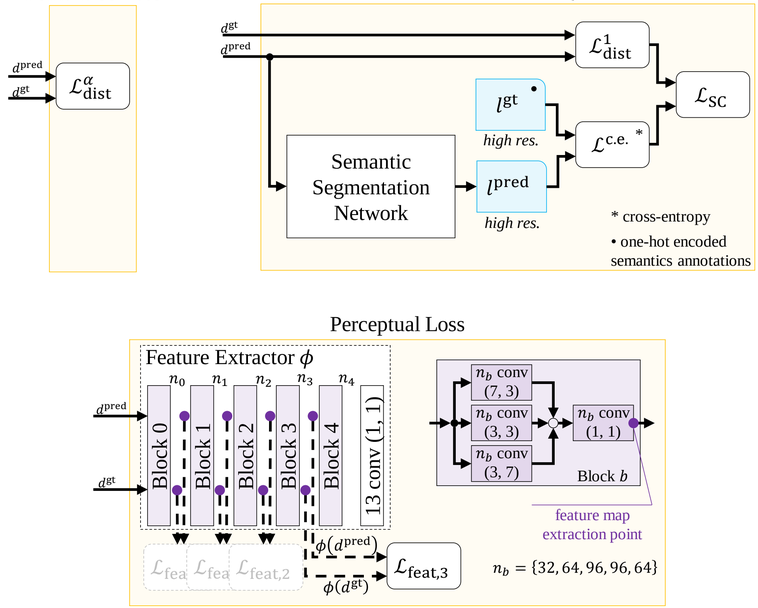
Motivation

LiDAR scanners are a **key enabler for autonomous driving**. They are required to have a very high resolution to **provide detailed information on the environment** and ensure a high detection performance. Generic point clouds are irregular and sparse, thus increasing the point density of the scans is not trivial. Given that our 3D point clouds exhibit a regular structure, we can use cylindrical 2D projections. Operating in 2D **preserves the regular structure in our data**, which is beneficial for downstream perception algorithms. Further, since constantly moving towards higher resolution sensors, our approach enables the **re-usage of recorded data** when moving towards a higher resolution LiDAR in the future.

Method



Overview on the proposed architecture: The upper figure shows the overall architecture. At training time, the input to the network is a down-sampled distance image of size $L/2 \times W$ with information about the missing measurements. The residual up-sampling network outputs an up-sampled distance image of size $L \times W$ with in-network up-scaling. Both distance images are inputs to the loss (yellow). The bottom figure shows the three different loss functions under consideration (only one is used at a time).



Modified point-wise loss:

$$\mathcal{L}^{\alpha}_{\text{dist}} = \frac{1}{\alpha\,|\mathcal{V}|} \sum_{(i,j)\in\mathcal{V}} \left| d^{\text{gt}}_{ij} - d^{\text{pred}}_{ij} \right|^{\alpha} \qquad \alpha = 1, 2$$

where $\alpha = 1$ describes the mean average error and $\alpha = 2$ describes the mean squared error with $\mathcal{V} = \{(i,j) \mid \text{reflection at } \theta_i,\, \varphi_j \text{ received}\}$

Perceptual loss:

$$\mathcal{L}_{\text{feat}} = \sum_{c,i,j} \left| \phi(\mathbf{d}^{\text{gt}})_{cij} - \phi(\mathbf{d}^{\text{pred}})_{cij} \right|$$

where c iterates over the different channels of the feature map

Multi-task semantic consistency (SC) loss function:

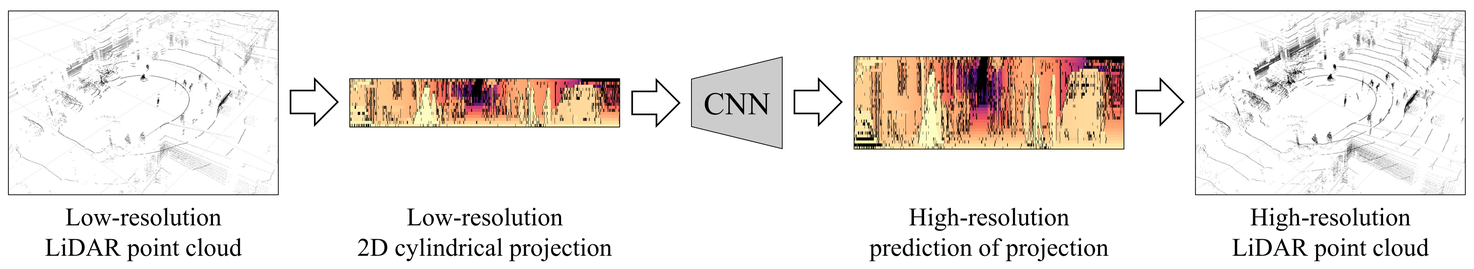
$$\mathcal{L}_{\text{SC}} = \frac{1}{2\sigma_r}\mathcal{L}^{1}_{\text{dist}} + \log\sigma_r + \frac{1}{\sigma_c}\mathcal{L}^{\text{cross-entropy}} + \log\sigma_c$$

where σ_r and σ_c are trainable variables that balance the relative weights of the two tasks

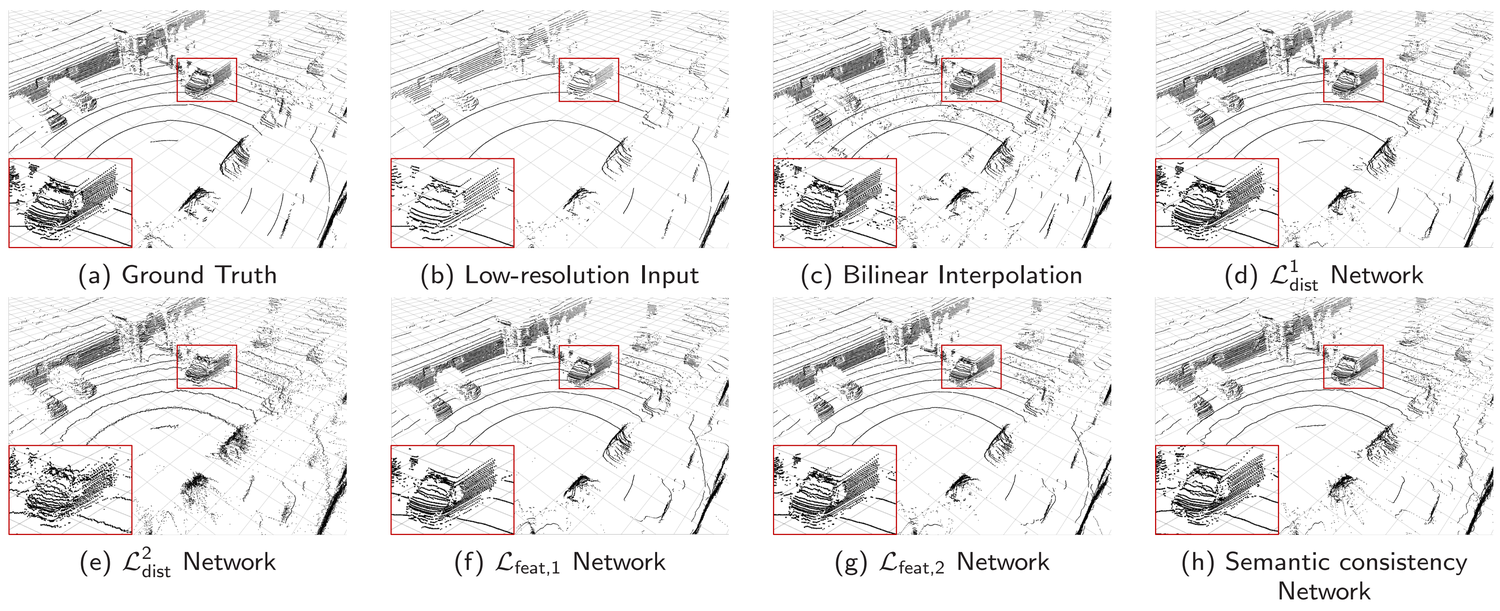
Contact

Larissa Triess
KIT and Daimler AG
Mail: larissa.triess@daimler.com

Overview



Examples



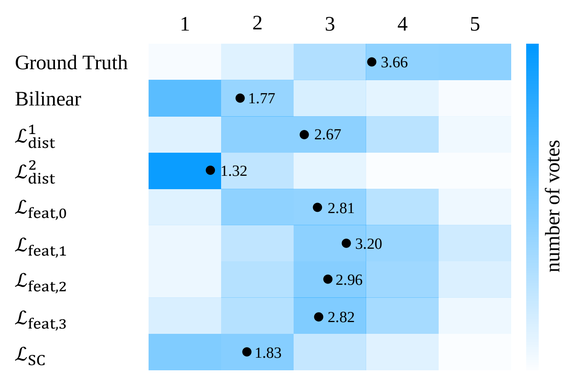
Synthesize (c) - (h) from (b) and compare to (a). Reconstruction quality mainly differs in high frequency perturbations in object boundaries, especially (e), and overall noise level, e.g. bilinear interpolation. The red rectangle enlarges the van visible in scene.

Quantitative Results

Networks	Semantics Dataset MSE [m]	Semantics Dataset MAE [m]	Semantics Dataset mIoU [%]	KITTI Raw MSE [m]	KITTI Raw MAE [m]
Ground truth	0.0	0.00	56.8	0.0	0.00
Bilinear	88.2	2.29	34.1	11.6	0.81
Bicubic	97.2	2.59	28.7	13.7	0.95
$\mathcal{L}^{1}_{dist}$	20.9	**0.68**	34.6	2.23	**0.21**
$\mathcal{L}^{2}_{dist}$	**17.6**	0.86	12.5	**1.95**	0.28
$\mathcal{L}_{feat,0}$	74.1	1.33	41.2	-	-
$\mathcal{L}_{feat,1}$	110.4	3.05	45.0	-	-
$\mathcal{L}_{feat,2}$	112.1	2.45	**49.4**	-	-
$\mathcal{L}_{feat,3}$	74.1	1.49	49.1	-	-
$\mathcal{L}_{SC}$	18.1	0.86	47.4	-	-

Qualitative Assessment

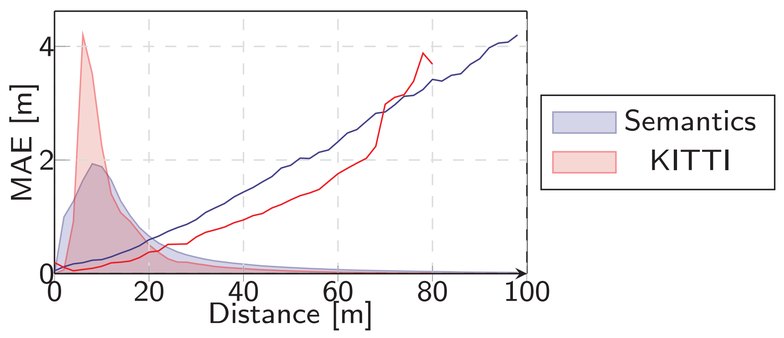
In the **mean optinion score (MOS)** survey we asked LiDAR experts to evaluate the *visual quality* of the point clouds generated by the proposed networks (1 - bad, 5 - excellent). Shown is the color-coded distribution (blue shades) of the MOS (circular marker) for ten randomly selected frames from the validation dataset.



Datasets

Two different datasets were used.

	Training	Validation	Test
Semantics [4]	344,027	73,487	137,682
KITTI	28,548	5,982	11,499



The plot shows the mean absolute error of the $\mathcal{L}^{1}_{dist}$ network as a function of the (ground truth) distance on the Semantics and KITTI datasets. The shaded areas depict the (normalize) distance distribution over each of the datasets.

Conclusion

Quantitative: choice of best performing model highly application-specific; different models excel in different evaluation metrics with respect to geometric and semantical accuracy

Qualitative: human subjects favored model variants involving perceptual loss based on visual realism as a performance criterion

Future work: design a single method that optimizes all our performance metrics

References

[1] L. Yu et al., "PU-Net: Point cloud upsampling network", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018

[2] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution", in *European Conference on Computer Vision*, 2016

[3] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017

[4] F. Piewak et al., "Boosting lidar-based semantic labeling by cross-modal training data generation", in *Computer Vision - ECCV 2018 Workshops*, 2018