Adversarial Semantic Scene Completion from a Single Depth Image

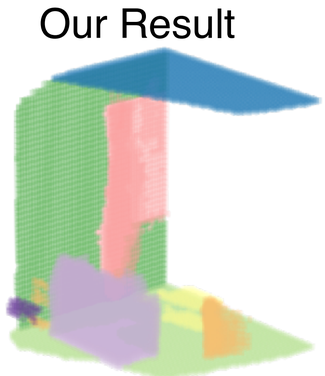
Yida Wang, David Joseph Tan, Nassir Navab, Federico Tombari
Technische Universität München

Overview

Depth Image

3D Completion + Semantic Labeling

Our Result



Empty
Ceiling
Floor
Wall
Window
Door
Chair
Bed
Sofa
Table
Furniture
Objects

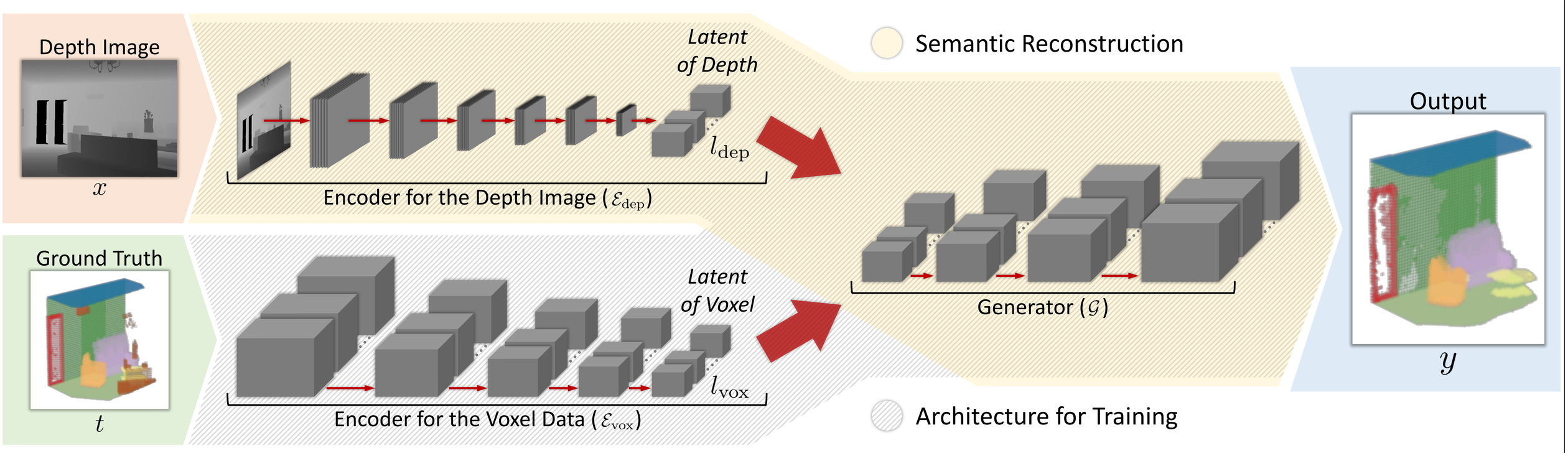
We introduce a direct reconstruction method to reconstruct from a 2.5D depth image to a 3D voxel data with both shape completion and semantic segmentation that relies on a deep architecture based on 3D VAE with an adversarial training to improve the performance of this task.

Related Works

Comparison between different methods targeted on 3D shape completion or semantic segmentation based on 2D or 2.5D informations. Our work does shape completion and semantic labelling based on a depth image.

Methods	Input	Output	Completion	Segmentation
3D-RecGAN++ [4]	TSDF Volumes	Volumetric Data	Yes	No
SSCNet[5]	TSDF Volumes	Volumetric Data	Yes	Yes
3D-R2N2[6]	RGB Image	Volumetric Data	Yes	No
Ours	Depth Image	Volumetric Data	Yes	Yes

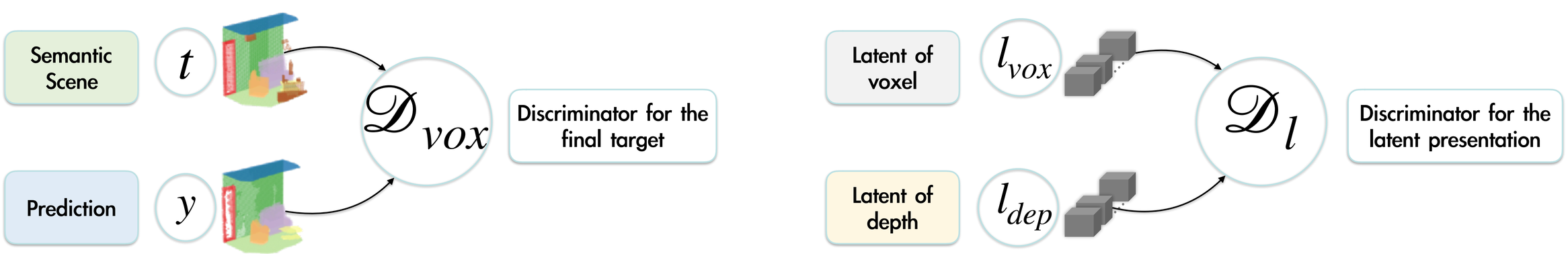
Our Method



We utilize the latent representation of 3D auto-encoder to help train a latent representation from a depth image. The 3D auto-encoder is removed after the parametric model is trained. This pipeline is optimized for the encoders for the depth image and the 3D volumetric data and the shared generator is also optimised during training.

Discriminators

To make the latent representation and the reconstructed 3D scene similar to each others, we apply two discriminators for both targets. In this manner, the latent representation of the depth produces the expected target more precisely compared to the latent representation of the ground truth volumetric data. Both of the discriminators are optimized to improve the ability in distinguishing the produced output from the given ground truth. Therefore, the discriminators help in fine-tuning the latent representation and the final output.



Quantitative Results

Semantic scene completion results on the SUNCG test set with depth map for IoU in percentage

	empty	ceil.	floor	wall	win.	door	chair	bed	sofa	table	furn.	objs.	Avg.
3D VAE [3]	49.3	26.1	33.2	29.7	14.4	4.6	0.7	16.4	13.9	0.0	0.0	0.0	30.8
3D-RecGAN++ [4]	49.3	32.6	37.7	36.0	23.6	13.6	8.7	20.3	16.7	9.6	0.2	3.6	36.1
Ours without Dl	49.6	42.0	35.9	44.8	28.5	25.5	15.4	28.6	20.1	21.5	11.5	6.5	42.7
Ours without Dvox	49.6	39.0	35.7	43.4	26.8	23.8	18.5	29.2	22.4	16.8	10.4	5.3	41.7
Ours (Proposed)	49.7	41.4	37.7	45.8	26.5	26.4	21.8	25.4	23.7	20.1	16.2	5.7	44.1

[1] A. Brock, T. Lim, J. M. Ritchie, and N. Weston. Generative and discriminative voxel modeling with convolutional neural networks. arXiv preprint arXiv:1608.04236, 2016.

[2] H. Fan, H. Su, and L. Guibas. A point set generation network for 3d object reconstruction from a single image. In Conference on Computer Vision and Pattern Recognition (CVPR), volume 38, 2017.

[3] A. Brock, T. Lim, J. M. Ritchie, and N. Weston. Generative and discriminative voxel modeling with convolutional neural networks. arXiv preprint arXiv:1608.04236, 2016.

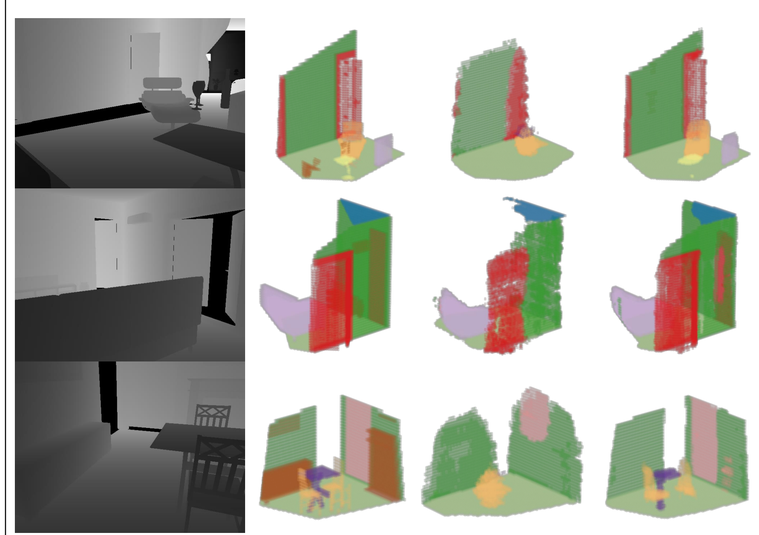
[4] B. Yang, S. Rosa, A. Markham, N. Trigoni, and H. Wen. 3d object dense reconstruction from a single depth view. arXiv preprint arXiv:1802.00411, 2018.

[5] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic scene completion from a single depth image. IEEE Conference on Computer Vision and Pat- tern Recognition, 2017.

[6] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d- r2n2: A unified approach for single and multi-view 3d object reconstruction. In Proceedings of the European Conference on Computer Vision (ECCV), 2016.

Qualitative Results

Depth Image | Ground Truth | 3D VAE [3] | Ours



We overcome the difficulty of semantically labeling small objects in 3D spaces and in sparsely completed surfaces.