Weakly-supervised 3D semantic scene reconstruction

Alexander Baumann[1] Sophia Wagner[1]
[1]Technical University of Munich

Problem Definition and Contribution

Goal: 3D semantic scene reconstruction of point cloud indoor scenes in a weakly-supervised fashion

Motivation:

- Applications to various tasks such as robot navigation and interior design
- Ground truth values for shape completion are not available in real-world point cloud scans as the full 3D geometry of objects is unknown due to occlusions, view constraints and weak illumination.

Key Contributions:

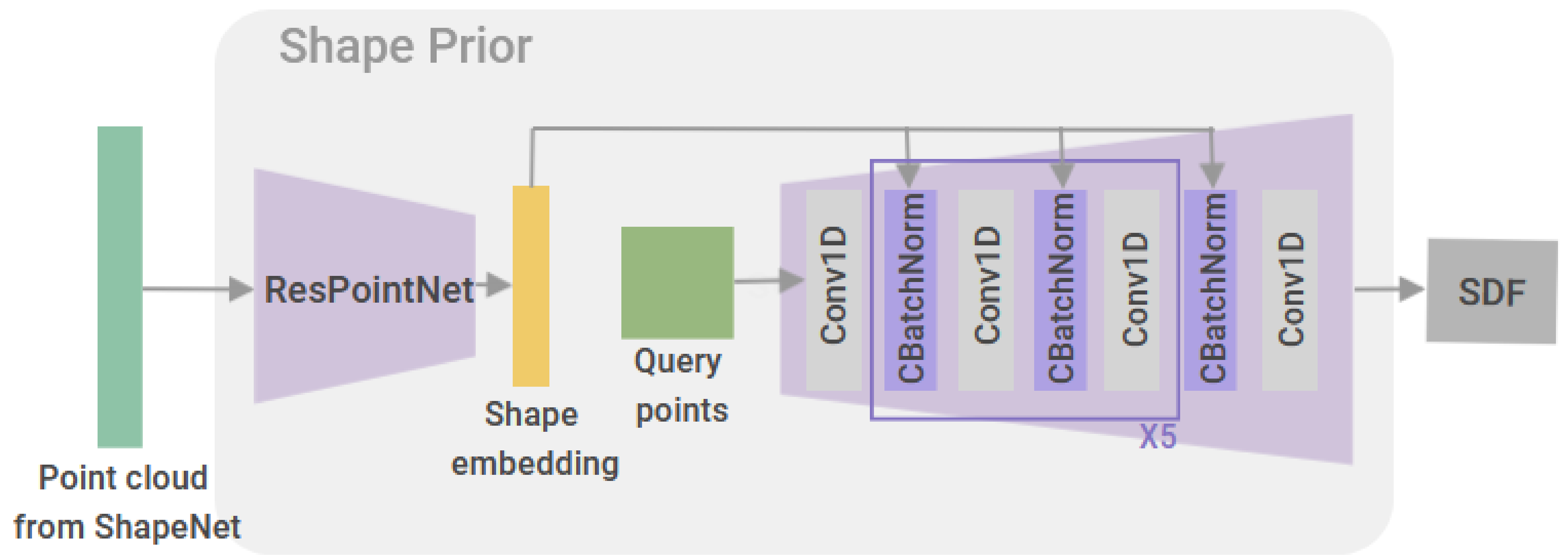
- A model for semantic instance reconstruction without using any supervision on the reconstruction of the shapes of real-world point cloud scans
- A shape prior which can predict the signed distance field directly from a given input point cloud, without voxelizing the scene

Method

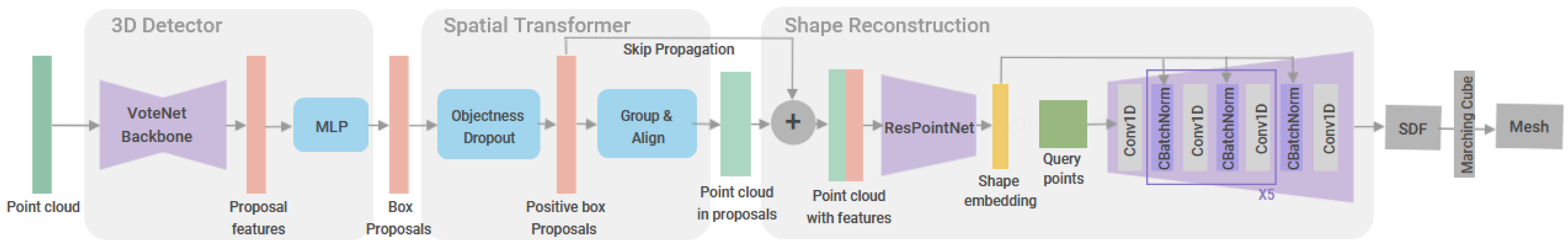
- Detect and semantically segment objects within 3D indoor scenes and group and align the corresponding point clouds into the canonical coordinate system
- Train a **shape prior** on the synthetic ShapeNet dataset[1] that generates meshes from a point cloud
- Integrate the shape prior into the full pipeline using different approaches:
  - by finetuning the **pretrained encoder** of the shape prior such that the resulting mesh is similar to objects of the same category of the ShapeNet data[1]
  - by training a new encoder that takes a point cloud as well as the predicted semantic label and 3D geometry information as input using a **skip propagation**. Again, the objective is that the resulting mesh is similar to objects of the same category of the ShapeNet data[1].

[1]Chang, Angel X., et al. "Shapenet", 2015

Architecture of the Shape Prior



Architecture of the full network



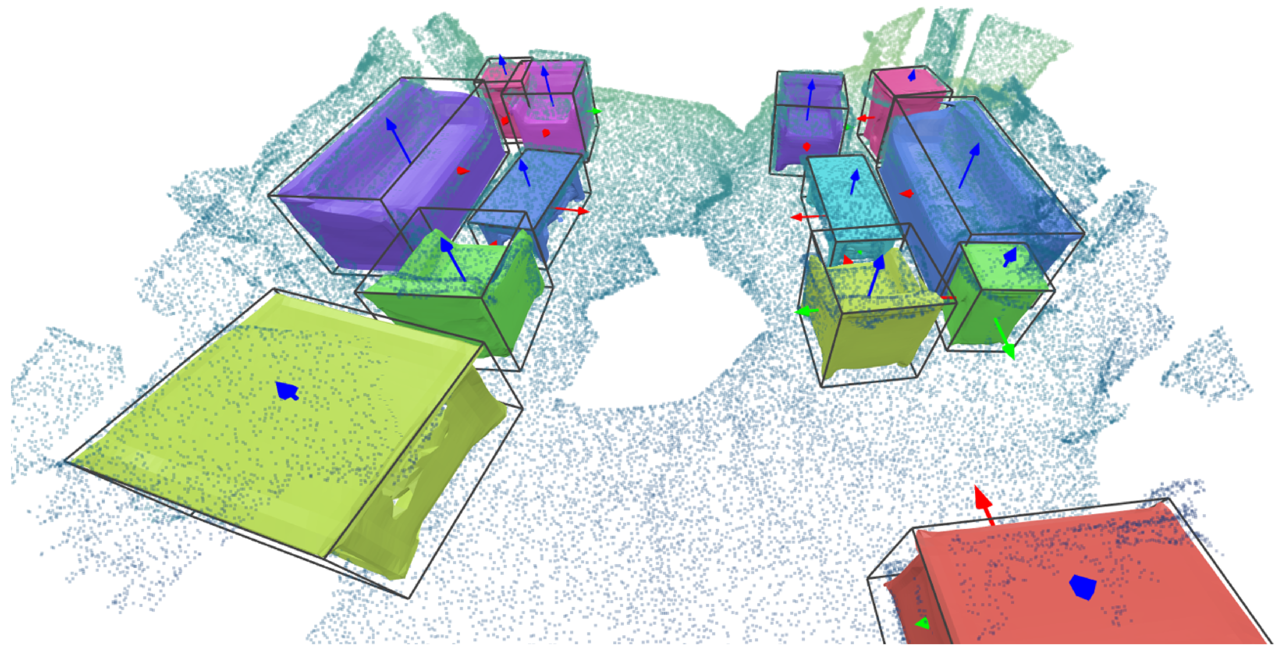
Results on 3D semantic scene reconstruction

- After having tried different losses and retrieval techniques, we deduced that an MSE-loss works best to retrieve either the mean shape embedding or a random embedding of the associating instance class.
- As can be seen in the table, we outperform the baseline in some categories even though no supervision is used for the shape reconstruction.

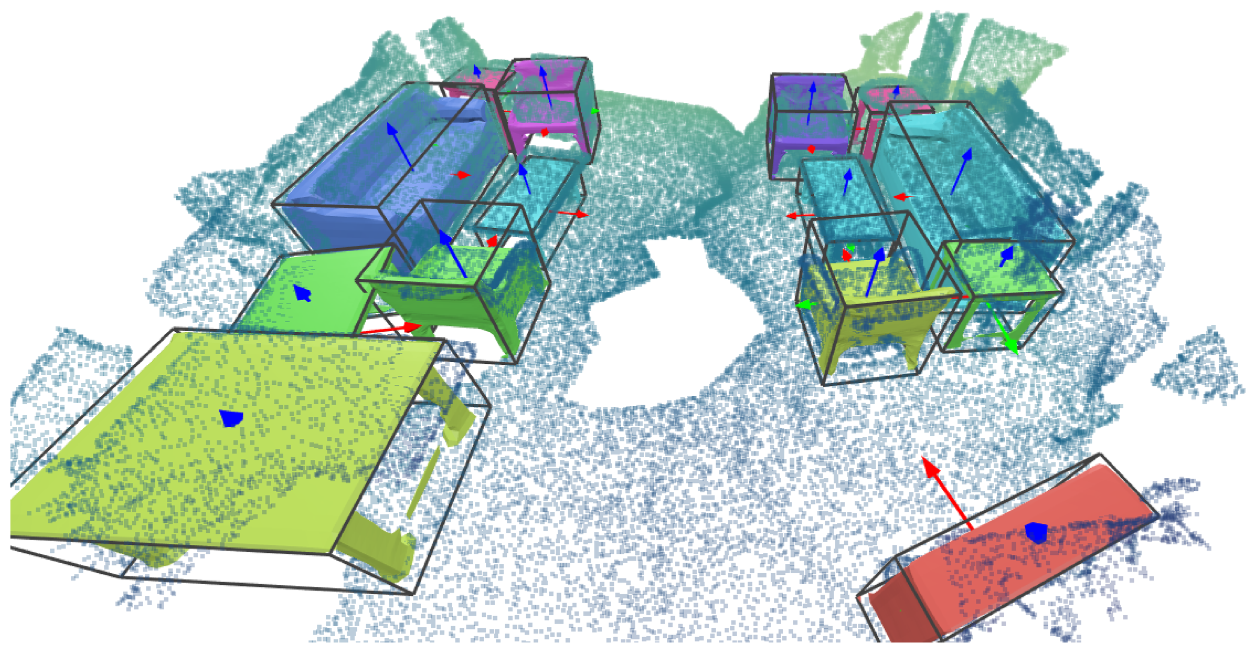
	table	chair	bookshelf	sofa	trash bin	cabinet	display	bathtub	overall
RfD-Net[2]	**0.729**	**0.126**	**0.507**	1.325	0.020	**0.345**	**0.014**	0.118	**0.411**
Skip propagation + mean embeddings	1.570	0.308	0.972	**1.134**	**0.012**	0.586	0.018	0.104	0.747
Skip propagation + random embeddings	1.587	0.296	0.696	1.380	0.024	0.650	0.019	**0.088**	0.731
Pretrained encoder + mean embeddings	1.494	0.266	0.764	1.492	0.021	0.453	0.024	0.126	0.690
Pretrained encoder + random embeddings	1.310	0.270	0.706	1.757	0.026	0.526	0.025	0.135	0.676

average Chamfer distances from the input point cloud to the corresponding extracted mesh, multiplied with 1000

[2]Nie, Yinyu, et al., "Rfd-net", 2021



prediction with RfD-Net



prediction with our model

Conclusion and summary

- We propose a novel approach for reconstructing the shapes without using any supervision on the given real-world dataset.
- Despite the advantage of ground-truth shapes of the corresponding supervised method, we can outperform this method in some of the semantic classes with respect to the chamfer distance.
- Adding more object classes into the training of the shape prior would result into more appropriate meshes on the real-world dataset.