Supplementary Material for Learning Priors for Semantic 3D Reconstruction

Ian Cherabier[1,⋆] Johannes L. Schönberger[1,*] Martin R. Oswald[1]
Marc Pollefeys[1,2] Andreas Geiger[1,3]

[1]ETH Zürich [2]Microsoft [3]MPI-IS and University of Tübingen

Abstract. In this supplementary document, we provide further insights and details about our novel multi-scale optimization scheme. In particular, we prove the adjointness of the transformation matrices of the primal and dual update steps of the unrolled optimization. Moreover, we detail the upsampling operators of the multi-scale scheme and show further examples of learned shape priors across different scale levels. Finally, we present additional experimental results which emphasize the fact that our method with unrolled optimization performs well when trained on little data, while traditional fully convolutional networks fail to reach comparable accuracy levels when trained with the same amount of data.

1 Multi-scale Optimization Scheme

This section provides additional details of our proposed multi-scale optimization scheme.

Derivation and Intuition. As described in the paper, the main energy in a single-scale setting can be numerically minimized using the first-order primal-dual (PD) algorithm in [1] with the following update steps:

$$\begin{aligned} &1.\ \nu^{t+1} = \nu^t + \sigma\left(\sum\nolimits_\ell \bar{u}_\ell^t - 1\right) && 3.\ u^{t+1} = \Pi_{[0,1]}\left[u^t + \tau(W^*\xi^{t+1} - f)\right] \\ &2.\ \xi^{t+1} = \Pi_{\|\cdot\|\leq 1}\left[\xi^t + \sigma W\bar{u}^t\right] && 4.\ \bar{u}^{t+1} = 2u^{t+1} - u^t \end{aligned} \tag{1}$$

Step 1 updates the Lagrangian dual variable which ensures the simplex constraint on the objective. Step 4 is an extrapolation step which speeds up the optimization and which is necessary to ensure convergence. The important steps of the algorithm are steps 2 and 3 which represent a (projected) gradient descent in the primal variable and a (projected) gradient ascent in the dual variable.

Remember, that W is a generalized learned gradient operator. Therefore, a good way to get intuition on the behaviour of W is to think in terms of classic gradient ascent (resp. descent) in step 2. (resp. step 3), without projection. That is, in step 2, the original gradient ascent scheme, without projection, was defined as

$$\xi^{t+1} = \xi^t + \sigma\nabla\bar{u}^t \tag{2}$$

⋆ These authors share first authorship.

In the multi-scale setting we have multiple versions of variables at different resolution scales s. The idea is to combine the gradient updates of the current resolution with the ones from the coarser scale with a linear combination:

$$\xi_s^{t+1} = \xi_s^t + \sigma\big(\nabla_s^s \bar{u}_s^t + U_{s+1}^s \nabla_{s+1}^s \bar{u}_{s+1}^t\big) \tag{3}$$

where ∇_s^s is the gradient operator on the current resolution, ∇_{s+1}^s is the gradient operator on the coarser scale and U_{s+1}^s is the upsampling operator in order to combine the two terms at the current scale level. This combined gradient ascent scheme then follows the deepest ascent directions of both resolution levels at the same time. Now, we can go back and replace the gradient operator with our proposed generalized matrix W and reintroduce the projection to arrive at the equation we presented in the paper:

$$\xi_s^{t+1} = \Pi_{\|\cdot\|\leq 1}\Big[\xi_s^t + \sigma\big(W_s^s \bar{u}_s^t + U_{s+1}^s W_{s+1}^s \bar{u}_{s+1}^t\big)\Big] \tag{4}$$

Here, it is important to note that we learn the matrices W implicitly also their scaling which defines the relative contribution of the two ascent steps. Without the learned weighting of these terms this gradient ascent scheme is unlikely to work. The derivation of the primal update step with the gradient descent is analogous.

The proposed multi-scale optimization incorporates information from coarser resolution levels (if available) in order to perform the update step. This leads to a weighted sum of gradient-based updates for the current and the next coarser resolution. In addition, the learned weights also regulate the step size of individual gradient descent and ascent steps which are crucial for algorithmic efficiency and to ensure convergence. An example of learned weights for different scale levels is given in Figure 1. Note that due to the learned weights, we cannot give any convergence guarantees for our multi-scale optimization. A necessary condition for the convergence of the primal-dual algorithm is however the adjointness of the primal and dual updates which we show in the following.

Adjointness of Update Steps in Primal-Dual Algorithm. In order to apply the primal-dual algorithm [1], the minimization problem over the voxel labeling u is first transformed into a saddle point problem by introducing a dual variable which results from replacing the norm in the TV-regularization by its Legendre-Fenchel dual. In a discretized setting, we obtain

$$\|Wu\|_1 = \max_{\|\xi\|_\infty \leq 1} \langle Wu, \xi \rangle \ . \tag{5}$$

The algorithm further uses the adjoint W^* of matrix W to transfer the differential operator from the primal to the dual variable:

$$\langle Wu, \xi \rangle = \langle u, W^* \xi \rangle \ . \tag{6}$$

This leads to the property that the primal and dual gradient terms in the update equations of the algorithm must be adjoint. For our novel multi-resolution

scheme, this property is not obvious and we therefore show the adjointness of the update steps in the following. The update steps for the multi-resolution optimization in the main paper are as follows:

$$\xi_s^{t+1} = \Pi_{\|\cdot\|\leq 1}\Big[\xi_s^t + \sigma\big(W_s^s \bar{u}_s^t + U_{s+1}^s W_{s+1}^s \bar{u}_{s+1}^t\big)\Big] \tag{7}$$

$$u_s^{t+1} = \Pi_{[0,1]}\Big[u_s^t + \tau\big(W_s^{s*}\xi_s^{t+1} + \overline{U}_{s+1}^s W_{s+1}^{s*}\xi_{s+1}^{t+1}\big) + \tau\left(\nu^{t+1} - f\right)\Big] \tag{8}$$

$$\bar{u}^{t+1} = 2u^{t+1} - u^t \tag{9}$$

where U_{s+1}^s is the upsampling operator from scale $s+1$ to scale s in the space of dual variables, and $\overline{U}_{s+1}^s$ is the upsampling operator in the space of primal variables. A necessary condition for viability of the primal-dual algorithm is therefore the following proposition.

Proposition 1 (Adjointness of Multi-scale Primal-dual Updates). *The transformation matrices of the multi-resolution primal dual updates*

$$W_s^s \bar{u}_s^t + U_{s+1}^s W_{s+1}^s \bar{u}_{s+1}^t \quad \textit{and} \quad W_s^{s*}\xi_s^{t+1} + \overline{U}_{s+1}^s W_{s+1}^{s*}\xi_{s+1}^{t+1} \tag{10}$$

are adjoint operators.

Proof. We first rewrite the dual update step as a single matrix vector multiplication by stacking the matrices and vectors

$$\Big\langle \quad \begin{pmatrix} W_s^s & U_{s+1}^s W_{s+1}^s \end{pmatrix}\begin{pmatrix} \bar{u}_s^t & \bar{u}_{s+1}^t \end{pmatrix}^T \;,\; \begin{pmatrix} I_s & U_{s+1}^s \end{pmatrix}\begin{pmatrix} \xi_s^{t+1} & \xi_{s+1}^{t+1} \end{pmatrix}^T \Big\rangle \tag{11}$$

$$= \Big\langle \quad \begin{pmatrix} I_s & U_{s+1}^s \end{pmatrix}\begin{pmatrix} W_s^s & W_{s+1}^s \end{pmatrix}\begin{pmatrix} \bar{u}_s^t & \bar{u}_{s+1}^t \end{pmatrix}^T \;,\; \begin{pmatrix} I_s & U_{s+1}^s \end{pmatrix}\begin{pmatrix} \xi_s^{t+1} & \xi_{s+1}^{t+1} \end{pmatrix}^T \Big\rangle \tag{12}$$

$$= \Big\langle \quad \begin{pmatrix} W_s^s & W_{s+1}^s \end{pmatrix}\begin{pmatrix} \bar{u}_s^t & \bar{u}_{s+1}^t \end{pmatrix}^T \;,\; \begin{pmatrix} I_s & U_{s+1}^s \end{pmatrix}^* \begin{pmatrix} I_s & U_{s+1}^s \end{pmatrix}\begin{pmatrix} \xi_s^{t+1} & \xi_{s+1}^{t+1} \end{pmatrix}^T \Big\rangle \tag{13}$$

$$= \Big\langle \quad \begin{pmatrix} W_s^s & W_{s+1}^s \end{pmatrix}\begin{pmatrix} \bar{u}_s^t & \bar{u}_{s+1}^t \end{pmatrix}^T \;,\; \begin{pmatrix} I_s & I_{s+1} \end{pmatrix}\begin{pmatrix} \xi_s^{t+1} & \xi_{s+1}^{t+1} \end{pmatrix}^T \Big\rangle \tag{14}$$

$$= \Big\langle \quad \begin{pmatrix} \bar{u}_s^t & \bar{u}_{s+1}^t \end{pmatrix}^T \;,\; \begin{pmatrix} W_s^{s*} & W_{s+1}^{s*} \end{pmatrix}\begin{pmatrix} \xi_s^{t+1} & \xi_{s+1}^{t+1} \end{pmatrix}^T \Big\rangle \tag{15}$$

where I_s is the identity at scale s. The transitions from (13) to (14) is due to the fact that since our upsampling uses a copy operator to go from scale $s+1$ to scale s, its adjoint is the corresponding downsampling operator. We can conclude

with a similar reasoning:

$$\left\langle \begin{pmatrix} \bar{u}_s^t & \bar{u}_{s+1}^t \end{pmatrix}^T, \begin{pmatrix} W_s^{s*} & W_{s+1}^{s*} \end{pmatrix} \begin{pmatrix} \xi_s^{t+1} & \xi_{s+1}^{t+1} \end{pmatrix}^T \right\rangle \tag{16}$$

$$= \left\langle \begin{pmatrix} \bar{u}_s^t & \bar{u}_{s+1}^t \end{pmatrix}^T, \begin{pmatrix} I_s & I_{s+1} \end{pmatrix} \begin{pmatrix} W_s^{s*} & W_{s+1}^{s*} \end{pmatrix} \begin{pmatrix} \xi_s^{t+1} & \xi_{s+1}^{t+1} \end{pmatrix}^T \right\rangle \tag{17}$$

$$= \left\langle \begin{pmatrix} \bar{u}_s^t & \bar{u}_{s+1}^t \end{pmatrix}^T, \begin{pmatrix} I_s & \overline{U}_{s+1}^s \end{pmatrix}^* \begin{pmatrix} I_s & \overline{U}_{s+1}^s \end{pmatrix} \begin{pmatrix} W_s^{s*} & W_{s+1}^{s*} \end{pmatrix} \begin{pmatrix} \xi_s^{t+1} & \xi_{s+1}^{t+1} \end{pmatrix}^T \right\rangle \tag{18}$$

$$= \left\langle \begin{pmatrix} I_s & \overline{U}_{s+1}^s \end{pmatrix} \begin{pmatrix} \bar{u}_s^t & \bar{u}_{s+1}^t \end{pmatrix}^T, \begin{pmatrix} I_s & \overline{U}_{s+1}^s \end{pmatrix} \begin{pmatrix} W_s^{s*} & W_{s+1}^{s*} \end{pmatrix} \begin{pmatrix} \xi_s^{t+1} & \xi_{s+1}^{t+1} \end{pmatrix}^T \right\rangle \tag{19}$$

$$= \left\langle \begin{pmatrix} I_s & \overline{U}_{s+1}^s \end{pmatrix} \begin{pmatrix} \bar{u}_s^t & \bar{u}_{s+1}^t \end{pmatrix}^T, \begin{pmatrix} W_s^{s*} & \overline{U}_{s+1}^s W_{s+1}^{s*} \end{pmatrix} \begin{pmatrix} \xi_s^{t+1} & \xi_{s+1}^{t+1} \end{pmatrix}^T \right\rangle \tag{20}$$

This proves the proposition. □

Upsampling Operators. Note that the upsampling operators $U_{s+1}^s, \overline{U}_{s+1}^s$ operate on different dimensions and spaces, respectively in the dual space and in the primal space. As shown above, they are also adjoint operators and, in our setting, the adjointness corresponds to a matrix transpose. For upsampling the primal and dual variables, we duplicate the values of coarser voxels on the finer resolution level. Let the weight matrix be $W_s^s \in \mathbb{R}^{2^{-s}M \times 2^{-s}N}$ and $W_{s+1}^s \in \mathbb{R}^{2^{-(s+1)}M \times 2^{-(s+1)}N}$, then the upsampling matrix $U_{s+1}^s \in \mathbb{R}^{2^{-s}M \times 2^{-(s+1)}M}$ is binary and contains ones only when voxel indices j in the coarse grid correspond to voxel indices i in the finer grid:

$$(U_{s+1}^s)_{ij} = \begin{cases} 1 & \text{if } i = \text{upSampleIdx}(j) \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

For a given voxel index of a coarser resolution level, the function upSampleIdx$(\cdot)$ returns the corresponding index in the upsampled, next finer voxel grid.

The second upsampling operator $\overline{U}_{s+1}^s$ is similarly defined with the difference being that $\overline{U}_{s+1}^s \in \mathbb{R}^{2^{-s}N \times 2^{-(s+1)}N}$.

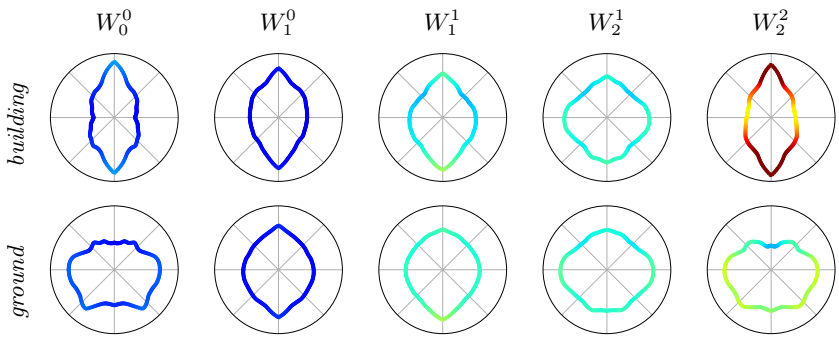

Fig. 1: **Learned multi-scale 3D shape priors** for label transitions to *free space*. The influence of the current and successive finer resolution to the update of the primal and dual variables has similar magnitude (*e.g.*, similar magnitude of W_0^0 and W_1^0 contributing to update of u_s^{t+1} and ξ_s^{t+1}), while the overall magnitude of the weights between different levels are increasing with higher scales to account for the different resolutions (*e.g.*, $W_0^0 < W_0^1 < W_0^2$). Note that the step sizes of the primal-dual algorithm and the weighting between data cost and regularization are factored into W. In contrast to manually choosing these hyper-parameters as in traditional approaches, our method learns the balancing of these parameters automatically from data.

2 Additional 2D Segmentation Results

In Figure 3 we present additional qualitative results of the 2D experiments we used to study and assess the behavior of our model. In this series of examples we can see the various benefits that our approach to segmentation gains over traditional TV-L1 approaches. The first is the capacity to propagate information through large region with missing data (see for instance example 3). A second useful benefit is that our method is able to adapt the segmentation to the labels. In traditional TV-L1, the regularizer looks for the minimal boundary between different labels. This may lead to unrealistic segmentations. Our approach on the other hand incorporates shape priors in its regularizer, which leads to more accurate segmentations (see,*e.g.*, image 2). Finally, our method learns sophisticated label interactions and ordering such as the fact that a roof can only exist above a building (see image 6).

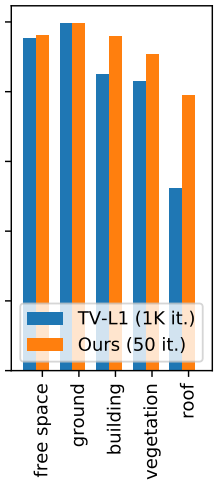

Fig. 2: **Per-label accuracies** for the semantic segmentations on synthetic images.

Figure 2 depicts a per-label accuracy plot for the 2D semantic segmentation experiments on synthetic data. It demonstrates that our learned semantic priors are much more useful for scene completion and denoising than the isotropic TV prior which does account for geometric nor semantic pixel neighborhood relationships.

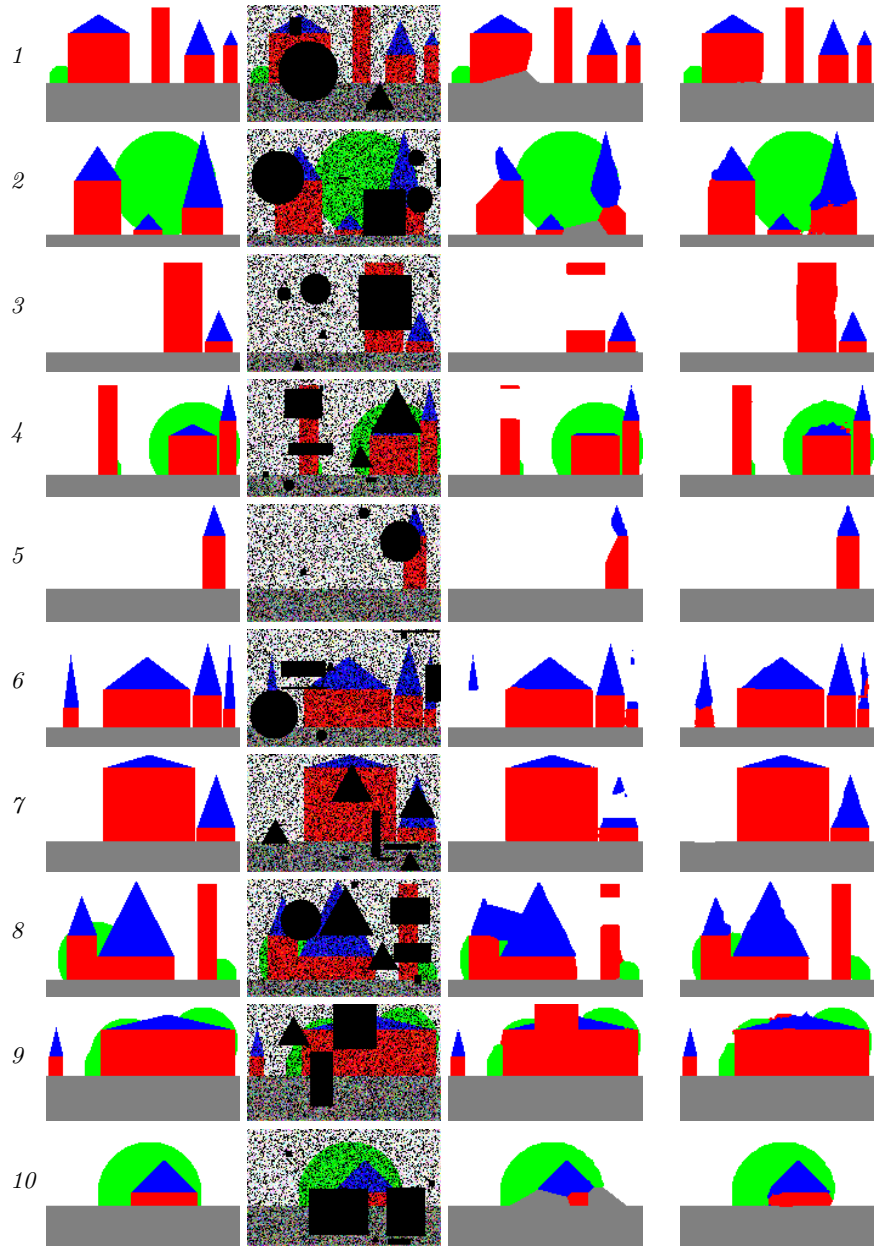

Fig. 3: **2D semantic segmentations on synthetic images.** Our method is capable of filling-in missing regions for shape completion. In case of image 6, our method is able to hallucinate a *building* below the observation of a roof, even though all data supporting the building is missing. This demonstrates the strong label interactions that our method is able to learn (*roof* can only be located on top of *building*).

3 Additional 3D Reconstruction Results

3.1 Quantitative Results

We present additional quantitative results for 3D reconstruction in Tab. 1. More specifically, we introduce some other baseline against which we compare our method. First of all, we present accuracy results for a solution extracted from the input datacost (entry *Input data* in Tab. 1. We then trained our method without unrolled optimization. This corresponds to a classic fully convolutional network with alternating convolution layers and ReLU activation layers. We trained this network on 300 scenes, but also on only 5 scenes which were selected by choosing the 5 training scenes with the largest variety of present semantic labels. The results are presented at the entry *Ours-5 (0 it.)* and *Ours-300 (0 it.)* in Tab. 1. We can see that this fully convolutional approach obtains good results when trained on 300 scenes, with an overall accuracy of 97.3%, though lower than our approach trained on 300 scenes. What is more interesting, is that the resutls are considerably degraded when trained on only 5 scenes, with overall accuracy of 65.1%, where our approach still reaches an overall accuracy of 96.7%. This illustrates how our method with unrolled optimization can learn strong models from little data.

Methods	Overall	Freespace	Occupied	Semantic
Input data	59.8	39.1	99.7	68.4
TV-L1 (50 it.)	92.8	71.0	91.4	87.8
TV-L1 (500 it.)	95.8	86.4	92.3	88.5
C2F (50 it.)	21.0	26.7	**99.9**	31.4
Ours-5 (0 it.)	65.1	98.0	83.5	76.2
Ours-5 (50 it.)	96.7	95.8	93.9	86.4
Ours-300 (0 it.)	97.3	97.6	92.3	90.2
Ours-300 (50 it.)	**98.7**	**98.6**	94.4	**91.5**

Table 1: **3D Reconstruction accuracy for ScanNet** [2]. The table reproduces the table from Fig. 7 in the paper, with the additional results for "Ours-5 (0 it.)" corresponding to our method without unrolled optimization trained on 5 scenes only. These results show that the variational unrolling significantly helps the network to generalize from very small amounts of training data and generally improves the overall performance.

3.2 Qualitative Results

In Figures 4, 5, 6, and 7 we show additional 3D reconstruction results for the ScanNet [2] dataset and outdoor scenes used in [3]. The reconstructions of our method are often more complete than the ground truth scene as our model learns

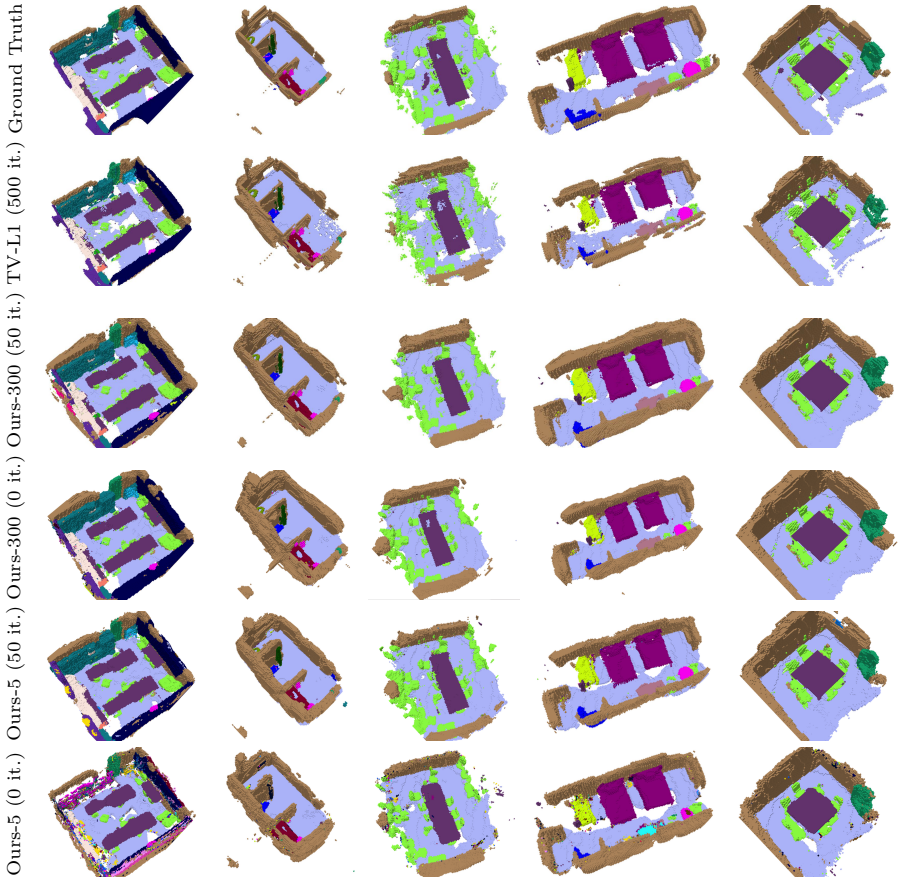

Fig. 4: **3D reconstruction results for ScanNet** [2] for different types of scenes and methods.

how to interpolate missing data, resulting in plausible and visually pleasing semantic 3D reconstruction results compared to the incomplete scans provided as ground truth. While classic fully convolutional network (our method without unrolled optimization) work well when trained on all 300 scenes, a significant loss in performance can be observed when trained on only 5 scenes. In Figure 7, we compare results of our approach with and without unrolled optimization, and see that the later helps with removing strong artifacts.

3.3 Application on SUNCG Data

We tested our method on the more challenging task of single view reconstruction by applying it to the synthetic indoor dataset SUNCG [4]. We wish first to stress that [4] solve a different problem than ours: they use a non-semantic TSDF as input and tune their approach for single view reconstruction. Our approach on

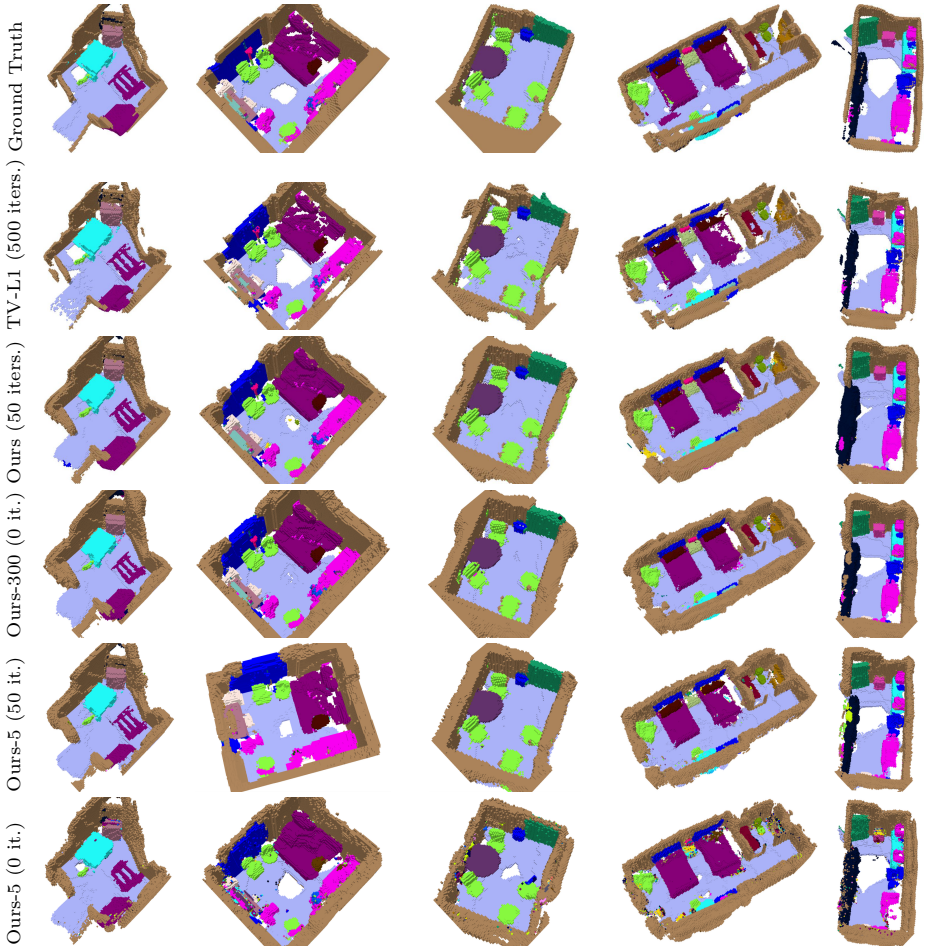

Fig. 5: **3D reconstruction results for ScanNet** [2] for different types of scenes and methods.

the other hand requires a multi-label input datacost, and is inherently adapted to a multi-view setting as it works by denoising and propagating data, rather than hallucinating new structures. Since we have neither tuned nor designed our method for such applications, this explains that the results are worse. We trained our multi-scale approach for 50 iterations and 3 scales on 500 views from SUNCG for 1770 epochs, and tested on 74 views. Results are presented in table 2.

The results support our intuition that our method is better suited for multi-view 3D reconstruction. We can see especially that it is more difficult to obtain good semantic accuracies.

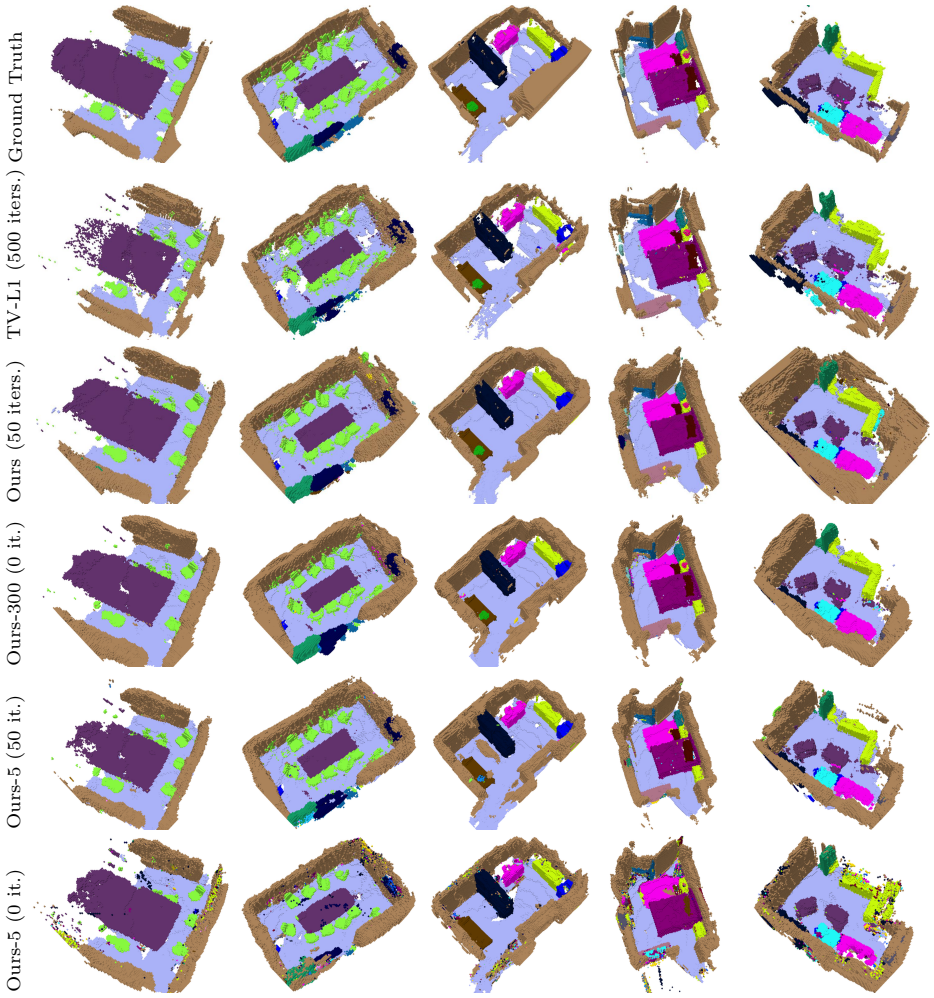

Fig. 6: **3D reconstruction results for ScanNet** [2] for different types of scenes and methods.

Table 2: **Accuracies of our network on SUNCG data [4]**

Method	Overall	Freespace	Occupied	Semantic
Ours-500 (50it - 3 level)	82.2	84.0	90.1	70.5

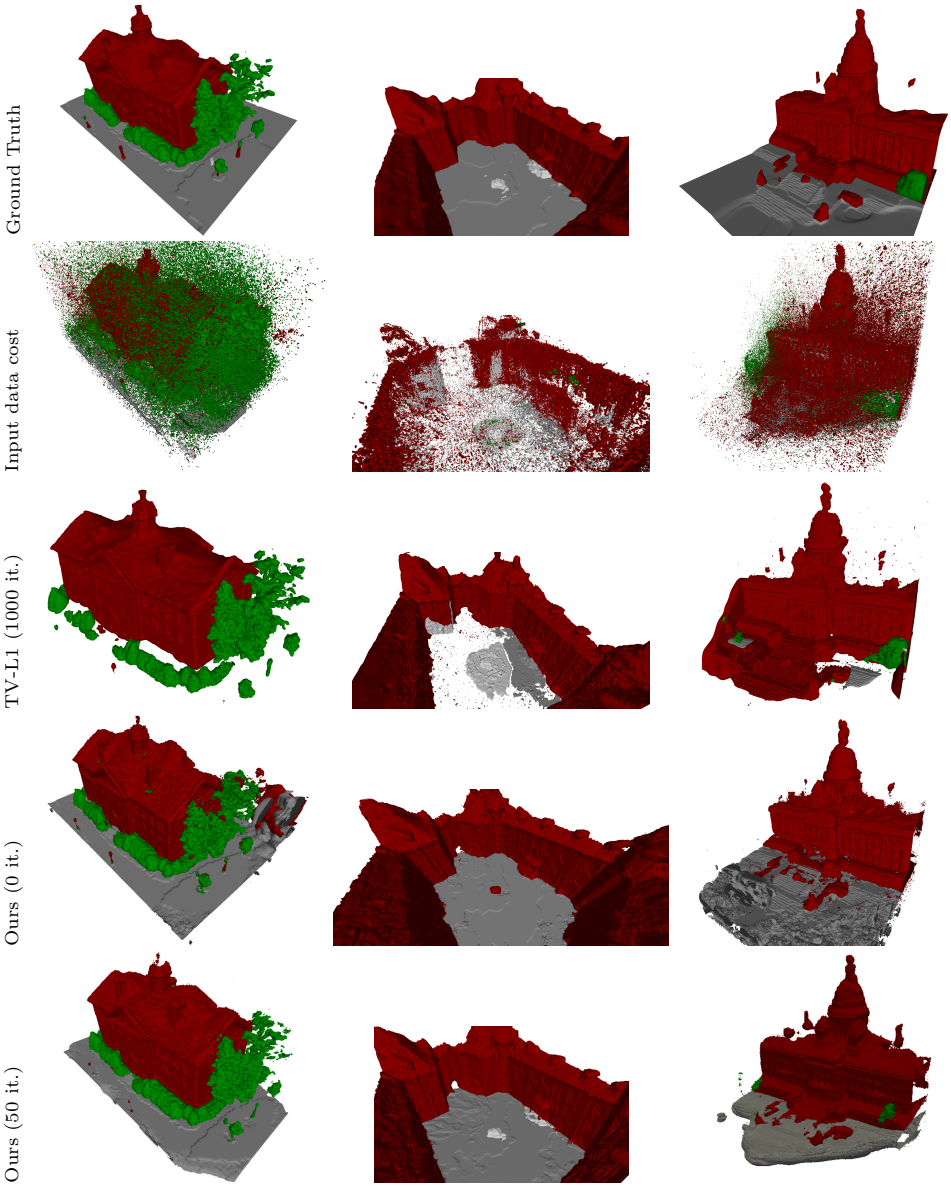

Fig. 7: **3D reconstruction results for Outdoor scenes.** We use the results from [3] as a groundtruth, and learn how to reproduce them. The results for TV-L1 with 1000 iterations correspond to convergence case. These results show that the absence of shape priors lead to incorrect reconstruction, such as the absence of ground. We can see that the results of our method without unrolled optimization can achieve good denoising, but with strong remaining artifacts for the left and right scene, or mislabeling in the case of the middle scene. Our approach with unrolled optimization overcomes these difficulties.

References

1. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of Mathematical Imaging and Vision (2011)
2. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proc. Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
3. Häne, C., Zach, C., Cohen, A., Angst, R., Pollefeys, M.: Joint 3d scene reconstruction and class segmentation. In: Proc. Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
4. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. Proc. Conference on Computer Vision and Pattern Recognition (CVPR) (2017)