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. We present a novel semantic 3D reconstruction framework which embeds variational regularization into a neural network. Our network performs a fixed number of unrolled multi-scale optimization iterations with shared interaction weights. In contrast to existing variational methods for semantic 3D reconstruction, our model is end-to-end trainable and captures more complex dependencies between the semantic labels and the 3D geometry. Compared to previous learning-based approaches to 3D reconstruction, we integrate powerful long-range dependencies using variational coarse-to-fine optimization. As a result, our network architecture requires only a moderate number of parameters while keeping a high level of expressiveness which enables learning from very little data. Experiments on real and synthetic datasets demonstrate that our network achieves higher accuracy compared to a purely variational approach while at the same time requiring two orders of magnitude less iterations to converge. Moreover, our approach handles ten times more semantic class labels using the same computational resources.

1 Introduction

Estimating 3D geometry from images is one of the long-standing goals in computer vision. Despite its long history, however, many problems remain unsolved. In particular, ambiguities arising from textureless or reflective regions, viewpoint changes, and image noise render the problem difficult. Powerful priors are therefore needed to robustly solve the task. One source of prior knowledge which can be exploited are semantics and their interaction with 3D geometry. Consider an urban scene, for example. While the ground is often flat and horizontal, building walls are mostly vertical and located on top of the ground. The availability of reliable semantic image classification methods has therefore recently driven the development of methods that jointly optimize geometry and semantics in 3D.

In their pioneering work, Häne *et al.* [10, 12, 13] proposed a method for joint volumetric 3D reconstruction and semantic segmentation using depth maps and semantic segmentations as input. They formulate the task as a variational multi-label problem, where each voxel is labeled by either one of the semantic classes or free space. Wulff shapes [28] serve as convex anisotropic regularizers, modeling the relationship between any two neighboring voxel labels. While impressive semantic reconstruction results have been demonstrated, the priors used

*These authors share first authorship.

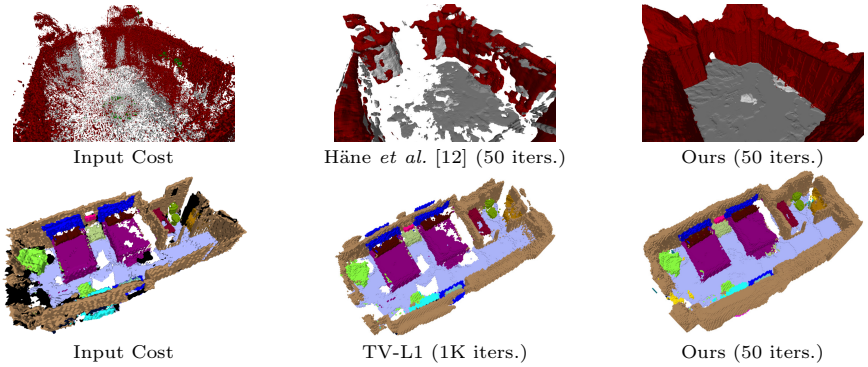


Fig. 1: **Semantic 3D reconstruction results.** We learn semantic and geometric neighborhood statistics to handle large amounts of noise, outliers, and missing data. Compared to traditional TV-L1 and the state of the art [12], our approach requires significantly less iterations and memory. Besides, it handles much larger label sets.

are hand-tuned and very simplistic, thus not able to fully capture the complex semantic and geometric dependencies of our 3D world. Furthermore, inference in those models requires thousands of iterations for convergence, limiting the applicability of these methods.

This work revisits the problem of jointly estimating geometry and semantics in a multi-view 3D reconstruction setting as shown in Fig. 1. Our approach combines the advantages of classical variational approaches [10, 12, 13] with recent advances in deep learning [32, 39], resulting in a method that is simple, generic, and substantially more scalable than previous solutions. In addition, our approach allows for automatically learning 3D representations from much fewer training data than existing learning-based solutions. As a result, our approach runs orders of magnitude faster than variational methods while producing better reconstructions. Moreover, memory requirements are significantly reduced allowing for larger label spaces. In summary, we make the following **contributions:**

- We present a novel framework for multi-view semantic 3D reconstruction which unifies the advantages of variational methods with those of deep neural networks, resulting in a simple, generic, and powerful model.
- We propose a multi-scale optimization strategy which accelerates inference, increases the receptive field, and allows long-distance information propagation.
- Compared to existing variational reconstruction methods [13], our approach learns semantic and geometric relationships end-to-end from data. Compared to fully convolutional architectures, our model is lightweight and can be trained from as little as five scenes without overfitting. Besides, formerly required manual and scene-dependent parameter tuning is no longer necessary and all meta-parameters, such as step sizes, are learned implicitly.
- Our experiments demonstrate that our method is able to achieve high quality results with only 50 unrolled optimization iterations compared to several thousands of iterations using traditional variational optimization.

Method	training scenes	model complexity	model parameters	runtime	manual tuning	#labels	semantic interactions
Learned [6, 8, 9, 36, 38, 40]	> 5K	high	millions	minutes	•	> 40	multi-scale
Variational (TV) [19, 31, 41]	> 0	low	none	seconds	•	> 40	none
Variational (Wulff-shape) [5, 10, 12]	> 1	moderate	hundreds	hours	•	< 10	single-scale
Learned-Variational [Ours]	**> 5**	**low**	**thousands**	**seconds**	•	**> 40**	**multi-scale**

Table 1: **Qualitative comparison of semantic reconstruction methods.** Quantities are approximate and categorized into positive, neutral, negative.

2 Related Work

Our work builds on a variety of computer vision and machine learning works. This section and Table 1 provide an overview of the most relevant prior works.

Semantic 3D Reconstruction. Ladicky *et al.* [22] presented a model for joint semantic segmentation and stereo matching. They considered simple height-above-ground properties as constraints between semantics and 3D geometry. Kim *et al.* [17] proposed a conditional random field (CRF) model for labeling the 3D voxel space based on a single RGB-D image and solved the CRF using graph cuts. Joint volumetric 3D reconstruction and semantic segmentation in a multi-view setting has been tackled by Häne *et al.* [12, 13] using variational optimization. Extensions to this seminal work consider object-class specific shape priors [10,23], scalable data-adaptive data structures [1], or larger semantic label spaces [5]. Kundu *et al.* [21] define a conditional random field to jointly infer semantics and occupancy from monocular video sequences.

A common drawback of these methods is that employed priors are either hand-crafted or not rich enough to capture the complex relationships of our 3D world. We propose to combine the advantages of variational semantic multi-view reconstruction with deep learning in an end-to-end trainable model. This leads to more accurate results and faster runtime as hyperparameters, such as the step size, are learned during training. Furthermore, we propose a novel multi-scale optimization scheme which allows to quickly propagate information across large distances and effectively increases the receptive field of the regularizer.

Variational Regularization. Variational energy minimization methods led to great advances when dealing with noise and missing information. A variety of regularizers have been studied in the literature [2–4, 27, 28, 42] in the context of different vision problems. Although these regularizers haven proven effective for low-level vision problems [3, 35] and 3D surface reconstruction [12, 19, 31, 41], they are limited in their expressiveness and do not fully capture the statistics of the underlying problem. In this paper, we propose a more expressive variational regularizer which jointly reasons at multiple scales and can be learned from data.

Learned Regularization. Several works combine the benefits of variational inference and deep learning. Early approaches combine proceed in a sequential manner by either learning the data costs for subsequent energy minimization [43] or by further regularizing the network output [14]. In contrast, several very re-

cent works integrate variational regularization directly into neural networks and apply them to 2D image processing tasks, including depth super-resolution [32], denoising [18, 25, 39], deblurring [18], stereo matching [39] and image segmentation [30]. Typically, the individual optimization steps are unrolled and embedded as layers into a neural network. Our work builds upon these ideas and tailors them to the multi-view semantic 3D reconstruction problem using a novel multi-scale neural network architecture for joint geometric and semantic reasoning.

Learned Shape Priors. Recently, deep learning based approaches have been proposed for depth map fusion [15], 3D object recognition [16, 24], or 3D shape completion [6, 8, 9, 36, 38, 40] using dense voxel grids as input. As all these approaches rely on generic 3D convolutional neural network architectures, they require a very large number of parameters and enormous amounts of training data. In contrast, our approach is more light-weight as it explicitly incorporates structural constraints via unrolled variational inference, therefore limiting the number of parameters needed. Although there are recent efforts to change the spatial scalability of these approaches using data-adaptive structures [11,33,34,37], current results are mostly limited to single objects or simple scenes and consider relatively small resolutions. However, none of these works have considered the semantic multi-view 3D reconstruction task which is the focus of this paper. Furthermore, our approach is fully convolutional and thus also scales to very large scenes.

3 Method

Using a generic 3D convolutional neural networks for semantic 3D reconstruction requires enormous amounts of memory and training data. In this paper, we therefore propose a more light-weight alternative which embeds a multi-label optimization task into the layers of a semantic 3D reconstruction network. We first introduce our multi-scale network architecture in Section 3.1, followed by a detailed description of the embedded variational problem in Section 3.2, and a description of the loss function we use for training the model in Section 3.3.

3.1 Network Architecture

The proposed network architecture for semantic 3D reconstruction is illustrated in Fig. 2. The input to our network is a set of semantically labeled depth maps aggregated into a 3D volume of truncated signed distance functions (TSDFs). More specifically, we follow [12] and accumulate per label evidence, *e.g.*, using depth maps from stereo and corresponding semantic image segmentations. As in traditional TSDF fusion, we trace rays from every pixel in each depth map to determine which voxels are occupied or empty. However, instead of using a fixed additive cost, we scale it using the semantic scores at the corresponding pixel. The output of our network is a volumetric semantic 3D reconstruction, where every voxel has one of the semantic class labels or the free space label.

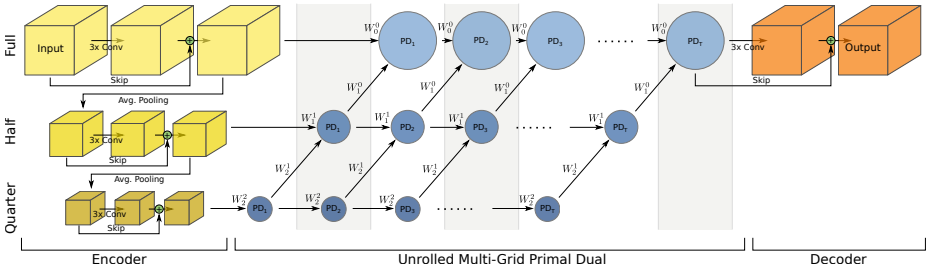


Fig. 2: **Proposed network architecture.** While the boxes represent data entities, the blue circles represent concurrent primal-dual (PD) processing steps with the iteration number as subscript. The weights W_i^j indicate the information flow (adjoint, primal and dual variables are omitted for brevity). The graph shows an example of our multi-scale optimization for three scales, however, their number is flexible.

Our network comprises three components (see Fig. 2): an encoder (yellow), the unrolled primal dual optimization layers (blue), and a decoder (orange). Our method reasons at multiple scales which allows for (i) modeling semantic interactions at different scales and (ii) propagating information quickly over larger distances during inference, *e.g.*, to complete missing data. We found that (i) results in higher accuracy while (ii) leads to much faster convergence compared to standard solvers [12, 42]. We now describe the three network components on a high level before providing a detailed derivation in Section 3.2.

Data Cost Encoder. At every voxel, the data cost is encoded by TSDFs computed via fused depth maps (*e.g.*, from stereo or Kinect) and semantic scene segmentations (*e.g.*, obtained from a semantic segmentation algorithm). In the first stage of our network, we pre-process this input using a shallow multi-scale neural network with 3 layers. The encoder serves several purposes: first, it normalizes the influence of the different semantic classes with respect to each other and the data term as a whole. Second, it helps in reducing low-level noise in the input. Finally, our multi-scale optimization requires down-sampling of the data cost which we learn automatically using separate encoders per scale. More concretely, starting at the highest resolution, we process the input with a residual unit that has two pairs of convolution-ReLU operations followed by a final convolution without activation. The encoded input is then down-sampled to the next scale using average pooling, followed by the next encoding stage.

Unrolled Multi-Grid Primal Dual. Instead of processing the input with a high-capacity 3D convolutional neural network, we propose to exploit variational optimization for semantic 3D reconstruction as a light-weight regularizer in our model. The advantage of such a regularizer is that it requires relatively few parameters due to temporal weight sharing while being able to propagate information over large distances by unrolling the algorithm for a fixed number of iterations and propagating information across multiple scales. More specifically, we unroll the iterations of the primal-dual (PD) algorithm of Pock and Chambolle [29], tailored to the multi-label semantic 3D reconstruction task, and

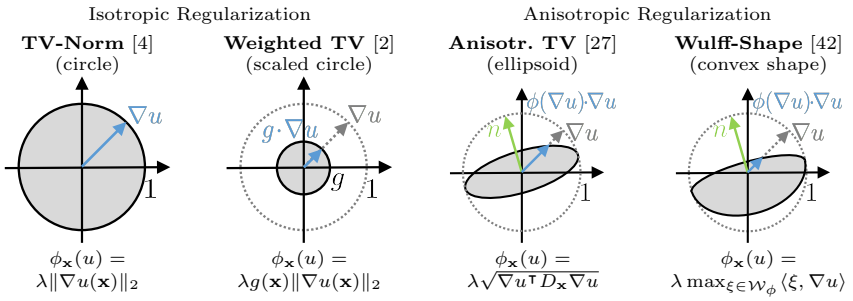


Fig. 3: **Overview of hand-crafted regularizers** that have been used in volumetric 3D reconstruction, e.g. weighted TV-Norm: [19, 41], Anisotropic TV: [20, 31], Wulff-shapes: [1, 12]. The polar plots show the smoothness cost $\phi_{\mathbf{x}}(\cdot)$ for different gradient directions ∇u. The right two cost functions are aligned to a given normal $\boldsymbol{n}$. We learn these regularization functions from the training data.

parameterize it by replacing the gradient operator with matrices which model the interaction of semantics and geometry at multiple scales for efficient label propagation. Each PD update equation defines a layer in the network, as illustrated with the blue circles in Fig. 2. To learn the parameters of the semantic label interactions and the hyper-parameters of the optimization algorithm, we back-propagate their gradients through the unrolled PD algorithm. A detailed derivation of our algorithm is presented in Section 3.2.

Probability Decoder. Similar to the proposed encoding stage, we also decode the obtained solution after the final PD iteration. The main goal here is to smooth and increase contrast, enabling stronger decisions on the final labeling and thereby improving accuracy. Our decoder takes the primal variable after the final iteration of the variational optimizer and feeds it into a residual unit with two pairs of convolution-ReLU operations followed by a final convolution with softmax activation for normalization.

3.2 Learning Variational Energy Minimization

This section describes the multi-grid primal-dual optimization algorithm which we leverage as light-weight, learned regularizer in our network. The traditional variational approach to volumetric 3D reconstruction [1, 10, 12, 13, 19, 20, 31] minimizes the energy

$$\underset{u}{\text{minimize}} \int_{\Omega} \Big(\underbrace{\phi_{\mathbf{x}}(u)}_{\text{regularization}} + \underbrace{fu}_{\text{data fidelity}} \Big) \, d\mathbf{x} \quad \text{subject to} \quad \forall \mathbf{x} \in \Omega : \sum\nolimits_{\ell} u_{\ell}(\mathbf{x}) = 1 \tag{1}$$

in order to find the best labeling $u : \Omega \rightarrow [0,1]^{|\mathcal{L}|}$ that assigns each point in space a probability for each label $\ell \in \mathcal{L}$. The constraint in (1) ensures normalized probabilities across all labels $\ell \in \mathcal{L}$ at every point $\mathbf{x} \in \Omega$. The data cost term

$f : \Omega \to \mathbb{R}^{|\mathcal{L}|}$ aggregates the noisy depth measurements of likely surface locations and is usually modeled as a truncated signed distance function (TSDF). To deal with noise, outliers, and missing data, a regularization term is typically added to the energy functional to obtain a smoother and more complete solution. The simplest choice for regularization is the total variation (TV) norm [2,4] $\phi_{\mathbf{x}}(u) = \lambda g(\mathbf{x}) \|\nabla u(\mathbf{x})\|_2$ which corresponds to minimizing the surface area of a 3D shape [19]. In most cases the weight function $g : \Omega \to \mathbb{R}_{\geq 0}$ encodes photoconsistency measures to align the surface with the input data. In many works this model has been extended to better deal with fine geometric details [20, 27, 31] or multiple semantic labels and directional statistical priors [12, 42]. Figure 3 provides an overview of various regularizers which have been proposed for 3D reconstruction. Notably, all these regularizers are convex and a global minimizer of Eq. (1) can be computed efficiently [3]. These hand-crafted regularizers are usually designed for tractability during optimization, but are not powerful enough to represent the true statistics of the underlying problem [18].

Proposed Energy. To overcome the limitations of hand-crafted regularizers, we follow Vogel and Pock [39] and generalize the gradient operator in the regularizer to the general matrix W, *i.e.* $\phi_{\mathbf{x}}(u) = \|Wu\|_2$. Since we are interested in modeling the complete space of directional and semantic interactions in the 3D multi-label setting, we choose to use a 6-dimensional matrix $W \in \mathbb{R}^{2\times2\times2\times|\mathcal{L}|\times|\mathcal{L}|\times3}$ for our task. This matrix computes gradients using forward-backward differences (modeled by $2 \times 2 \times 2$) and can represent higher-order interactions between any combination of semantic labels (modeled by $|\mathcal{L}| \times |\mathcal{L}|$) in any spatial direction (modeled by last dimension 3). For $W = \nabla$, we obtain a standard TV regularizer. Note that in contrast to the Wulff shapes used by [12], representing W directly leads to a large reduction in the number of parameters and consequently in memory as evidenced by our experimental evaluation. In this work, we aim to learn the weights of this matrix jointly with the other network parameters considering the following energy minimization problem:

$$\underset{u}{\text{minimize}} \quad \int_{\Omega} \left(\|Wu\|_2 + fu \right) d\mathbf{x} \quad \text{subject to} \quad \forall \mathbf{x} \in \Omega : \sum\nolimits_{\ell} u_{\ell}(\mathbf{x}) = 1 \tag{2}$$

Optimization. To minimize the convex energy in Eq. (2), we use a first-order primal-dual (PD) algorithm [3] for which the problem is first transformed into a saddle point problem. We introduce the dual variable ξ to replace the TV-norm with its conjugate. We also relax the constraints in Eq. (1) by introducing the Lagrangian variable ν. Then, the corresponding discretized saddle point energy

$$\underset{u}{\text{minimize}} \max_{\|\xi\|_{\infty} \leq 1} \langle Wu, \xi \rangle + \langle f, u \rangle + \nu \left(\sum\nolimits_{\ell} u_{\ell} - 1 \right) \tag{3}$$

can be minimized using the update equations

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

at time t with a total of T iterations, W^* the adjoint of W, step sizes τ and σ and projections $\Pi_{[0,1]}$ and $\Pi_{\|\cdot\|\leq 1}$, see [3]. Note that the operations $W\bar{u}$ and $W^*\xi$ convolve the kernel W with the variables ξ and $\bar{u}$. This enables efficient integration of these operations into a CNN with shared weights across the primal and dual updates and across the different iterations of the algorithm. We embed this algorithm into our network architecture by unrolling it for a fixed number of iterations. The input to the unrolled PD network is the pre-processed data cost term f provided by the encoder and the output is the optimized primal variable u which is passed to the decoder for post-processing.

Optimization Unrolling. One pass on the updates in Eq. (4) corresponds to one PD iteration. Similar to [32], we unroll the PD algorithm for a fixed number of iterations. Each PD update equation defines a layer in the network, as illustrated with the blue circles in Fig. 2. This unrolled PD algorithm constitutes the core of the network that we use to learn the label interactions represented by W. Note that the step sizes σ and τ that appear in Eq. (4) influence the speed of convergence of the PD algorithm. These parameters are typically selected manually or by preconditioning [29]. In this work, we learn the step sizes automatically by factoring them into W and thereby eliminating them from the update equations, contributing to fast convergence of the proposed algorithm.

Multi-Scale Optimization. In the algorithm discussed above, information only propagates between neighboring voxels, generally resulting in slow convergence of the optimization. Therefore, label interactions are relatively low-level and cannot capture more complex statistics arising at larger scales. While it is easy to enlarge the spatial extent of the matrix W, a drawback of naïvely increasing W is the cubic increase in the number of parameters which slows down training and makes the model prone to overfitting. Hence, we consider an alternative in this paper: instead of increasing the size of W, we simultaneously consider the scene at multiple scales.

More specifically, at each PD iteration, information is passed from the lower to the higher scales, as shown in Fig. 2. This enables long-range propagation of information and recovery of fine details while at the same time allowing faster back-propagation of gradients during training. Besides, inference runs in parallel at different scales, which, in practice, results in another speedup of the optimization as compared to traditional coarse-to-fine approaches, where the optimization must wait for coarser scales to converge. Note that even with different regularizer matrices W for each scale, the increase in the number of parameters is at most linear in the number of scale levels. Thus, the increase is sub-linear in the receptive field size compared to the cubic increase of the single-scale approach.

In our network, information is propagated via the matrix W. Thus, we lift our model to multiple scales by modifying update steps 2 and 3 in Eq. (4) to

$$\xi_s^{t+1} = \Pi_{\|\cdot\|\leq 1}\Big[\xi_s^t + \sigma\big(W_s^s \bar{u}_s^t + U_{s+1}^s W_{s+1}^s \bar{u}_{s+1}^t\big)\Big] \tag{5}$$

$$u_s^{t+1} = \Pi_{[0,1]}\Big[u_s^t + \tau\big(W_s^{s*}\xi_s^{t+1} + U_{s+1}^{s*} W_{s+1}^{s*}\xi_{s+1}^{t+1}\big) + \tau\left(\nu_s^{t+1} - f\right)\Big] \tag{6}$$

where s is one of S scale levels (lower level = higher resolution) and U^s_{s+1} upsamples from $s+1$ to s. W^s_s corresponds to the regularizer at level s, while W^s_{s+1} handles the transfer of information from level $s+1$ to the next finer level s.

3.3 Loss Function

We train the network architecture in Fig. 2 using supervised learning. Towards this goal, we define the training objective as the semantic reconstruction loss between our computed solution u and a given ground truth labeling $\hat{u}$. Typically, this loss is defined as the categorical cross entropy. However, several important modifications to the standard definition of this loss are necessary in practice as the ground truth is often not completely observed or labeled. We follow common practice and introduce a separate label $\tilde{\ell}$ for unlabeled regions. Unobserved regions are modeled by a uniform distribution $\mathcal{U}_\mathcal{L}$ in label space. To make the loss function agnostic to unobserved areas in the ground truth and to not penalize our solution in unlabeled regions, we use the following weighted loss function

$$H(u, \hat{u}) = -\int_\Omega w(\mathbf{x})\, u(\mathbf{x}) \log \hat{u}(\mathbf{x})\; d\mathbf{x} \tag{7}$$

$$w(\mathbf{x}) = \Delta_{KL}(\hat{u}(\mathbf{x}), \mathcal{U}_\mathcal{L}) \Delta_{KL}(\hat{u}(\mathbf{x}), \delta_{\tilde{\ell}}) \tag{8}$$

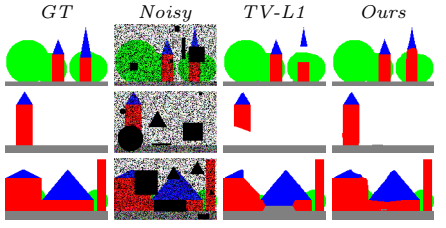
which returns zero if the ground truth at $\mathbf{x}$ is not unobserved or unkown. Here, Δ_{KL} denotes the KL-divergence. The first term measures the similarity between the ground truth and a uniform distribution and the second term the similarity to a Dirac distribution with center $\tilde{\ell}$. In case the ground truth matches exactly the uniform distribution or it is unlabeled with maximum certainty, this is equivalent to masking the loss as a hard constraint. However, as shown in the experiments, we generate ground truth using conventional regularization methods. As a result, it is beneficial to penalize using a soft constrained loss on the imperfectly labeled ground truth. Without the proposed weighting, the training would receive contradicting supervisory signals. Specifically, if the ground truth is incomplete for a specific class, the loss would encourage reconstruction in the observed areas whereas a potentially correct labeling in the unobserved parts would be inadvertently penalized.

4 Results

This section presents our results. We first analyze the memory and runtime complexity of our method wrt. to the state-of-the-art approach of Häne *et al.* [12]. Next, we empirically validate our approach in a controlled setting on a synthetic 2D toy dataset Finally, we present results on challenging indoor and outdoor semantic reconstruction tasks.

4.1 Memory and Runtime Complexity

One of the main advantages of our method over Häne *et al.* [12] is the significantly reduced memory complexity. While the approach of Häne *et al.* has a memory



T \ S	1	2	3	4	3+E	3+E+D	TV-L1
10	76.83	97.57	98.10	98.34	**99.37**	99.32	97.73
	38.58	82.11	87.43	88.74	94.94	**95.14**	79.50
20	90.76	98.26	98.85	98.86	99.38	**99.41**	98.40
	49.13	88.80	91.42	91.83	95.16	**95.23**	85.94
50	97.21	98.99	99.19	99.21	99.20	**99.38**	98.70
	74.36	91.56	91.42	93.20	93.57	**94.86**	88.31
1000	–	–	–	–	–	–	98.8
	–	–	–	–	–	–	89.2

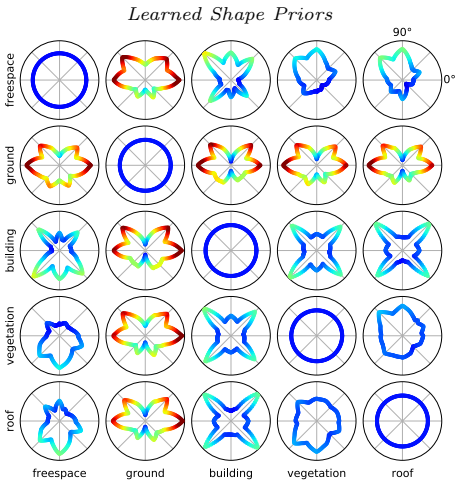


Fig. 4: **2D semantic segmentation on synthetic images. Top Left:** 3/1200 test scenes with ground truth (GT), noisy method input and the results of TV-L1 and ours in comparison. **Bottom Left:** Reconstruction accurcay for TV-L1 and our method using different numbers of iterations T (TV-L1 converges in 1000 iterations) and scales S. Accuracy over *all pixels* is shown in the first and accuracy only over *regions with missing data cost* in the second rows. **Right:** The plots show label transition costs between two labels depending on the surface normal. Ours learns more complex cost functions compared to the hand-crafted ones in Fig. 3. The cost functions have been rescaled for readability, with the magnitude encoded as color.

complexity of $(3+d)|\mathcal{L}|\cdot|\Omega| + (1+d)|\mathcal{L}|^2|\Omega|$, ours has a complexity of $(3+d)|\mathcal{L}|\cdot|\Omega| + 3\cdot 2^d|\mathcal{L}|^2$. Here, d is the dimension of Ω, and $|\mathcal{L}|$ and $|\Omega|$ the number of labels and voxels. Note that using additional scales in our approach only marginally increases the amount of memory, since each successive higher scale has 2^d fewer voxels. While their approach maintain dual variables for all label combinations at each location in the voxel grid, our approach shares this state for all locations. In practice, for a moderate scene size of $|\Omega| = 300^3$ (500^3) voxels with $|\mathcal{L}| = 40$ labels and single-precision floating point data, theirs has an intractable memory usage of around 668GB (3TB) versus a tractable 24GB (111GB) for ours. In addition to an improved memory complexity, our approach is much faster to compute. Compared to the costly calculation of Wulff shape projections, the convolution operations in our case are much cheaper to compute and, in practice, are implemented efficiently on GPUs. In summary, our proposed approach makes it tractable to perform joint semantic 3D reconstruction for both larger scenes and significantly more labels, as shown in the experiments.

4.2 Experiments on Synthetic 2D Data

Dataset. For validating our model, we created a simple 2D toy dataset with 5 labels, each defined by a color (white for *free space*, gray for *ground*, red for

building, blue for *roof* and green for *vegetation*). The scenes were generated with shapes like boxes, triangles, and circles, which were randomly positioned subject to interval bounds and ordering constraints, *e.g.*, roof on top of building and building on top of ground. We perturb the images with Gaussian noise and simulate missing data by removing large regions using random shapes (circle, square, triangle). Fig. 4 shows examples along with their degraded versions. We created 3000 images of size 160×96 for training and 1200 for testing, respectively. The data cost for label $\ell \in \mathcal{L}$ is defined as $f_\ell = \| I - c_\ell \|_2^2$ where I is the input image and c_ℓ is the color corresponding to label ℓ. For regions with missing pixels, which can only be filled by regularization, we use a uniform data cost.

Quantitative Evaluation. Using this dataset, we evaluate the benefit of the multi-scale approach as well as the feature encoding (E) and the probability decoding (D) networks. All networks are trained from random initialization with a batch size of 32. Fig. 4 (left) shows results on the test set with TV-L1 as a baseline. We show the accuracy computed on the whole image and only on the missing regions. The latter emphasizes the performance of the regularizer since in these regions, the data cost has no influence. Our approach consistently outperforms TV-L1, especially in the missing regions. This shows that our method learns more powerful regularizers, encoding statistics about geometry and semantics. Furthermore, increasing the number of scales and including encoding and decoding networks is beneficial.

Qualitative Evaluation. Fig. 4 (left) compares the segmentations from our full network ($T = 20$, $S = 3$) to those of TV-L1. While TV-L1 finds the (wrong) minimal surface solution, our network correctly fills in these regions and respects ordering constraints (*e.g.* building above ground).

Learned Priors. Our network learns costs at label transitions in a small 2^d neighborhood at every scale. This cost is influenced by the orientation of the transitions: vertical transitions between building and ground should be penalized more than horizontal transitions. Fig. 4 (right) plots the label transition costs against the surface normal for all label combinations. We see that the regularizer has the desired behavior in most cases, *e.g.* for building to ground transitions, we see that vertical transitions are penalized the most.

4.3 Experiments on Real 3D Data

We now use the best-performing architecture as determined in our 2D experiments and apply it to the 3D multi-label domain using two challenging datasets. We show that we can replicate hand-crafted Wulff shapes by learning from solutions produced by Häne *et al.* [12]. Using the learned weights, our approach produces equivalent results but two orders of magnitude faster, using only a fraction of the memory. Moreover, we apply our method to datasets with ten times more labels than can be handled by existing Wulff shape approaches.

Datasets. For all datasets, we assume gravity aligned inputs and use a standard multi-label TSDF for data cost aggregation [12]. For comparing against Häne *et al.* [13], we use their 3 outdoor scenes (*Castle*, *South Building*, *Providence*) with

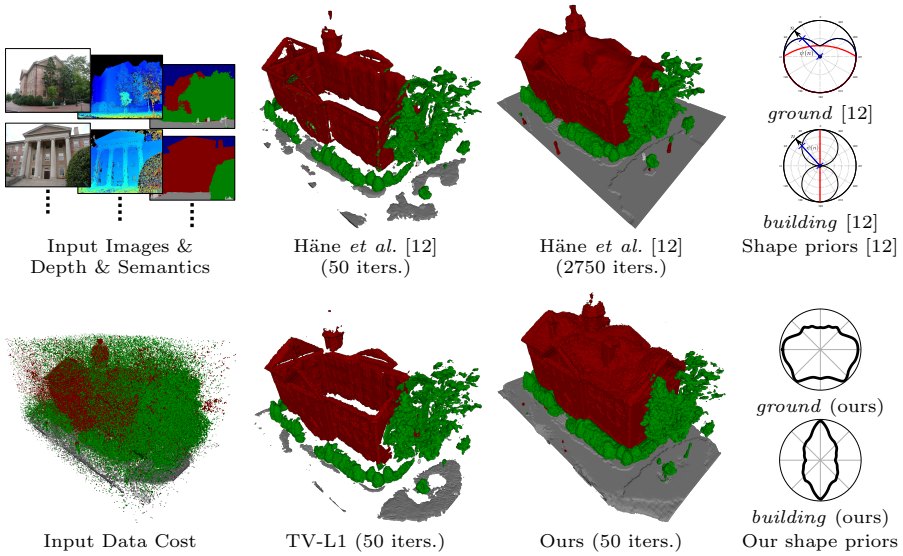


Fig. 5: **Semantic 3D reconstruction results. Left:** Input. **Middle:** Reconstruction results. Our method learns semantic and geometric neighborhood statistics to effectively handle large amounts of noise, outliers and missing data. Compared to traditional TV-L1 and the state-of-the-art [12], our approach requires significantly less iterations and memory. **Right:** Hand-crafted shape priors from Häne *et al.* [12] (top) vs. our learned shape priors (bottom).

5 labels (*freespace*, *ground*, *building*, *vegetation*, *unkown*). The largest scene has a size of around 300^3 voxels. In addition, we evaluate on the recently released ScanNet dataset [7], comprising 1513 scenes with fine-grain semantic labeling. We adopt the NYU [26] labeling with 40 classes. Using a voxel resolution of 5cm, the largest scenes have a size of around 400^3 voxels.

Training. Our network can optimize arbitrarily sized scenes both during inference and training, as our architecture is fully convolutional. However, due to the increased memory requirements during back-propagation and the computational benefits of batch processing in stochastic gradient descent, we train on fixed-size, random crops of dimension 32^3 with a batch size of 4 and a learning rate of 10^{-4}. We perform data augmentation by randomly rotating and flipping around the gravity axis. For all experiments, we unroll the PD algorithm for $T = 50$ iterations using $S = 3$ scales. As our network uses a few parameters as compared to pure learning approaches, overfitting is not a problem for our approach and training typically converges quickly after a few thousand mini batches.

Wulff Shape Comparison. First, we are interested in replacing the more complex and computationally costly Wulff shape approach [12] by learning from data produced by their method. Fig. 5 (right) shows the original Wulff shapes by Häne *et al.* next to our learned shapes at scale $s = 0$. The cost shape visualization is equivalent to the synthetic 2D experiments with the difference that here we

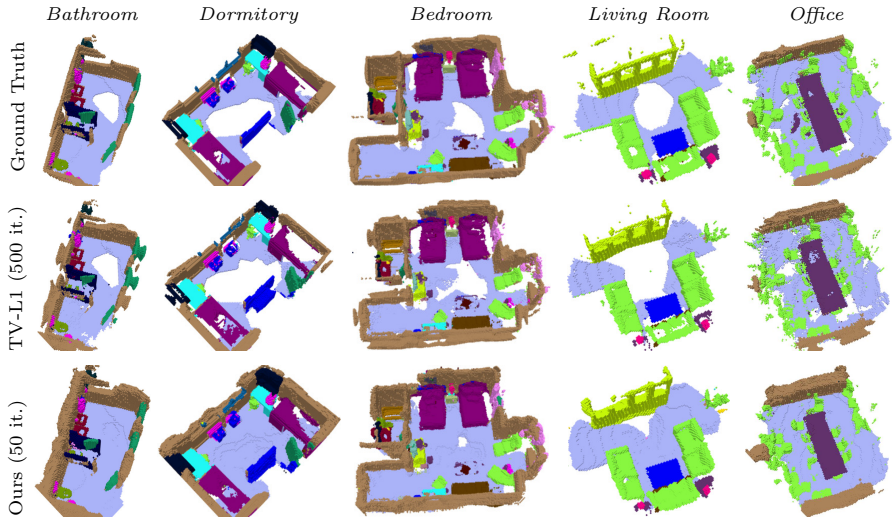


Fig. 6: **3D reconstruction results for ScanNet** [7] for different scenes and methods.

compute the average shape around the gravity axis. Our method meaningfully learns the hand-crafted shapes, demonstrating that we can replicate the more complex Wulff shape formulation. This is confirmed by a 98% per-class accuracy when evaluating our learned weights on the full scenes wrt. [12]. Figs. 1 and 5 show qualitative results for *Castle* and *South Building*. Moreover, our results are achieved after 50 iterations and 10 seconds while their approach requires 2750 iterations and around 4000 seconds to converge. Next, we demonstrate our method in a setting with an order of magnitude more class labels, which would be computationally intractable for their method [12].

Evaluation on ScanNet [7]. For ScanNet we re-integrate the provided depth maps and semantic segmentations using TSDF fusion based on the provided camera poses to establish voxelized ground truth. The resulting data costs provide very strong evidence and we thus use multi-label TV-L1 optimization with $W = \nabla$. For our evaluation, we also generate weak data costs by only integrating every 50th frame. The objective during training is to recover the high fidelity ground truth generated from the strong data cost using only the weak data cost as input. We train our network using 312 training scenes and evaluate their performance on 156 test scenes. Fig. 7 summarizes quantitative results for a reconstruction extracted from the input data cost, a multi-label TV-L1, a coarse-to-fine version of our network, a version of our network without variational regularization (0 iterations), and our proposed multi-scale architecture. Note that ours without regularization is a simple variant of approaches like SSCNet [36] or ScanComplete [6]. We draw the following conclusions: First, running TV-L1 for the same number of iterations as our method results in significantly worse results. Second, running TV-L1 for an order of magnitude more iterations until convergence still performs worse than our method. Third, a naïve coarse-to-fine

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 (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. 1 level)	98.7	98.6	94.4	91.5
Ours-300 (50 it. 3 levels)	**98.7**	**98.6**	**94.4**	**91.5**

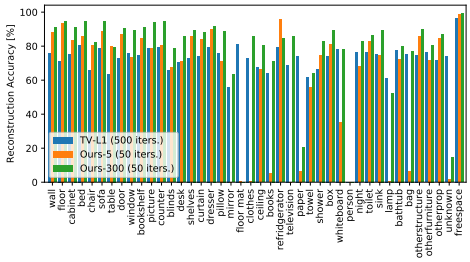


Fig. 7: **3D Reconstruction accuracy for ScanNet** [7]. **Left:** Reconstruction extracted from the input data, TV-L1 for 50 and for 500 iterations ($\widehat{=}$ converged), traditional coarse-to-fine network (C2F), our method with/without multi-scale scheme trained on 312 scenes, ours with multi-scale trained on a subset of only 5 scenes, ours without the unrolled optimization (0 iterations). **Right:** Per-label accuracies.

approach does not converge during training and produces bad reconstructions. Moreover, integrating multi-scale variational regularization into the network significantly improves the completeness of the results. Lastly, a version trained on only 5 scenes attains almost the same overall accuracy as a version trained on the full training dataset, indicating that our model can be trained with very little data. Furthermore, due to the few learned parameters in our network, we achieve the same accuracy for the training and test scenes which demonstrates the generalization power of our model. Figs. 1 and 6 show qualitative results for selected scenes. Surprisingly, our method sometimes produces results which are visually more pleasing than the ground truth used for training. We attribute this to the fact that our method can learn correct label interactions from all training data jointly and can then apply this knowledge to a single instance.

5 Conclusion

We presented a novel method for dense semantic 3D reconstruction. By incorporating variational regularization into a neural network, we can learn powerful semantic priors using a limited number of parameters. In stark contrast to purely learning based approaches, our method requires little training data and generalizes to new scenes without overfitting. The proposed multi-scale optimization jointly reasons about semantics and geometry at different scales and enables inference that is an order of magnitude more efficient than the state of the art. Experiments on synthetic and real data demonstrate the benefits wrt. accuracy, runtime, memory consumption, and algorithmic complexity.

Acknowledgements. This work received funding from the Horizon 2020 research and innovation programme under grant No. 637221 (Built2Spec), No. 688007 (TrimBot2020). This research was also supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number D17PC00280. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

1. Bláha, M., Vogel, C., Richard, A., Wegner, J.D., Pock, T., Schindler, K.: Large-scale semantic 3d reconstruction: an adaptive multi-resolution model for multi-class volumetric labeling. In: Proc. Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
2. Bresson, X., Esedoḡlu, S., Vandergheynst, P., Thiran, J.P., Osher, S.: Fast global minimization of the active contour/snake model. Journal of Mathematical Imaging and Vision (2007)
3. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of Mathematical Imaging and Vision (2011)
4. Chan, T., Esedoḡlu, S., Nikolova, M.: Algorithms for finding global minimizers of image segmentation and denoising models. SIAM Journal on Applied Mathematics (2006)
5. Cherabier, I., Häne, C., Oswald, M.R., Pollefeys, M.: Multi-label semantic 3d reconstruction using voxel blocks. In: International Conference on 3D Vision (3DV) (2016)
6. Dai, A., Ritchie, D., Bokeloh, M., Reed, S., Sturm, J., Nießner, M.: ScanComplete: Large-Scale Scene Completion and Semantic Segmentation for 3D Scans. In: Proc. Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
7. 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)
8. Dai, A., Qi, C.R., Nießner, M.: Shape completion using 3d-encoder-predictor cnns and shape synthesis. In: Proc. Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
9. Han, X., Li, Z., Huang, H., Kalogerakis, E., Yu, Y.: High-resolution shape completion using deep neural networks for global structure and local geometry inference. In: Proc. International Conference on Computer Vision (ICCV) (2017)
10. Häne, C., Savinov, N., Pollefeys, M.: Class specific 3d object shape priors using surface normals. In: Proc. Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
11. Häne, C., Tulsiani, S., Malik, J.: Hierarchical surface prediction for 3d object reconstruction (2017)
12. 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)
13. Häne, C., Zach, C., Cohen, A., Pollefeys, M.: Dense semantic 3d reconstruction. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2017)
14. Heber, S., Pock, T.: Convolutional networks for shape from light field. In: Proc. Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
15. Ji, M., Gall, J., Zheng, H., Liu, Y., Fang, L.: Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In: Proc. International Conference on Computer Vision (ICCV) (2017)
16. Kar, A., Tulsiani, S., Carreira, J., Malik, J.: Category-specific object reconstruction from a single image. In: Proc. Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
17. Kim, B.S., Kohli, P., Savarese, S.: 3d scene understanding by Voxel-CRF. In: Proc. International Conference on Computer Vision (ICCV) (2013)

18. Kobler, E., Klatzer, T., Hammernik, K., Pock, T.: Variational networks: Connecting variational methods and deep learning. In: Proc. German Conference on Pattern Recognition (GCPR) (2017)
19. Kolev, K., Klodt, M., Brox, T., Cremers, D.: Continuous global optimization in multiview 3d reconstruction. International Journal of Computer Vision (IJCV) (2009)
20. Kolev, K., Pock, T., Cremers, D.: Anisotropic minimal surfaces integrating photoconsistency and normal information for multiview stereo. In: Proc. European Conference on Computer Vision (ECCV) (2010)
21. Kundu, A., Li, Y., Dellaert, F., Li, F., Rehg, J.M.: Joint semantic segmentation and 3d reconstruction from monocular video. In: Proc. European Conference on Computer Vision (ECCV) (2014)
22. Ladický, L., Sturgess, P., Russell, C., Sengupta, S., Bastanlar, Y., Clocksin, W., Torr, P.H.S.: Joint Optimization for Object Class Segmentation and Dense Stereo Reconstruction. International Journal of Computer Vision (IJCV) (2012)
23. Mahabadi, R.K., Hane, C., Pollefeys, M.: Segment based 3d object shape priors. In: Proc. Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
24. Maturana, D., Scherer, S.: Voxnet: A 3d convolutional neural network for real-time object recognition. In: International Conference on Intelligent Robots and Systems (IROS) (2015)
25. Meinhardt, T., Möller, M., Hazirbas, C., Cremers, D.: Learning proximal operators: Using denoising networks for regularizing inverse imaging problems. In: Proc. International Conference on Computer Vision (ICCV) (2017)
26. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: Proc. European Conference on Computer Vision (ECCV) (2012)
27. Olsson, C., Byröd, M., Overgaard, N.C., Kahl, F.: Extending continuous cuts: Anisotropic metrics and expansion moves. In: Proc. International Conference on Computer Vision (ICCV) (2009). https://doi.org/10.1109/ICCV.2009.5459206, https://doi.org/10.1109/ICCV.2009.5459206
28. Osher, S.J., Esedolu, S.: Decomposition of images by the anisotropic rudin-osher-fatemi model. Communications on Pure and Applied Mathematics (2004)
29. Pock, T., Chambolle, A.: Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In: Proc. International Conference on Computer Vision (ICCV) (2011)
30. Ranftl, R., Pock, T.: A deep variational model for image segmentation. In: Jiang, X., Hornegger, J., Koch, R. (eds.) Pattern Recognition (2014)
31. Reinbacher, C., Pock, T., Bauer, C., Bischof, H.: Variational segmentation of elongated volumetric structures. In: Proc. Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
32. Riegler, G., Rüther, M., Bischof, H.: Atgv-net: Accurate depth super-resolution. In: Proc. European Conference on Computer Vision (ECCV) (2016)
33. Riegler, G., Ulusoy, A.O., Bischof, H., Geiger, A.: Octnetfusion: Learning depth fusion from data. In: International Conference on 3D Vision (3DV) (2017)
34. Riegler, G., Ulusoy, A.O., Geiger, A.: Octnet: Learning deep 3d representations at high resolutions. In: Proc. Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
35. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena (1992)

36. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.A.: Semantic scene completion from a single depth image. In: Proc. Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
37. Tatarchenko, M., Dosovitskiy, A., Brox, T.: Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In: Proc. International Conference on Computer Vision (ICCV) (2017)
38. Tulsiani, S., Zhou, T., Efros, A.A., Malik, J.: Multi-view supervision for single-view reconstruction via differentiable ray consistency. In: Proc. Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
39. Vogel, C., Pock, T.: A primal dual network for low-level vision problems. In: Proc. German Conference on Pattern Recognition (GCPR) (2017)
40. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: Proc. Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
41. Zach, C., Pock, T., Bischof, H.: A globally optimal algorithm for robust tv-l1 range image integration. In: Proc. International Conference on Computer Vision (ICCV) (2007)
42. Zach, C., Shan, L., Niethammer, M.: Globally optimal finsler active contours. In: Pattern Recognition (Proc. DAGM) (2009)
43. Zbontar, J., LeCun, Y.: Computing the stereo matching cost with a convolutional neural network. In: Proc. Conference on Computer Vision and Pattern Recognition (CVPR) (2015)