

# Where Does It End? – Reasoning About Hidden Surfaces by Object Intersection Constraints

Michael Strecke and Jörg Stückler

Embodied Vision Group, Max Planck Institute for Intelligent Systems, Tübingen

{mstrecke, jstueckler}@tue.mpg.de

## Abstract

*Dynamic scene understanding is an essential capability in robotics and VR/AR. In this paper we propose Co-Section, an optimization-based approach to 3D dynamic scene reconstruction, which infers hidden shape information from intersection constraints. An object-level dynamic SLAM frontend detects, segments, tracks and maps dynamic objects in the scene. Our optimization backend completes the shapes using hull and intersection constraints between the objects. In experiments, we demonstrate our approach on real and synthetic dynamic scene datasets. We also assess the shape completion performance of our method quantitatively. To the best of our knowledge, our approach is the first method to incorporate such physical plausibility constraints on object intersections for shape completion of dynamic objects in an energy minimization framework.*

## 1. Introduction

Humans have the remarkable ability to infer missing information using assumptions on the physical plausibility of scenes. For instance, if we see objects lying on a table, we immediately conclude that the shape of the objects is constrained by the table surface. We would be surprised if an object continues within the table.

In this paper we propose Co-Section, a novel approach to incorporate such constraints into a bottom-up 3D scene reconstruction method. Our method tracks and maps dynamic objects in the scenes. In dynamic scenes, however, we can use more than just the image measurements to reason about the object shapes: individual objects cannot penetrate each others while they move relative to each others and eventually are in contact. In contrast to a variety of learning-based approaches for shape completion [6, 18, 5, 24], we incorporate such physical plausibility as intersection constraints in an energy-minimization framework.

Our formulation optimizes the implicit surfaces of the objects as a signed distance functions from oriented points.

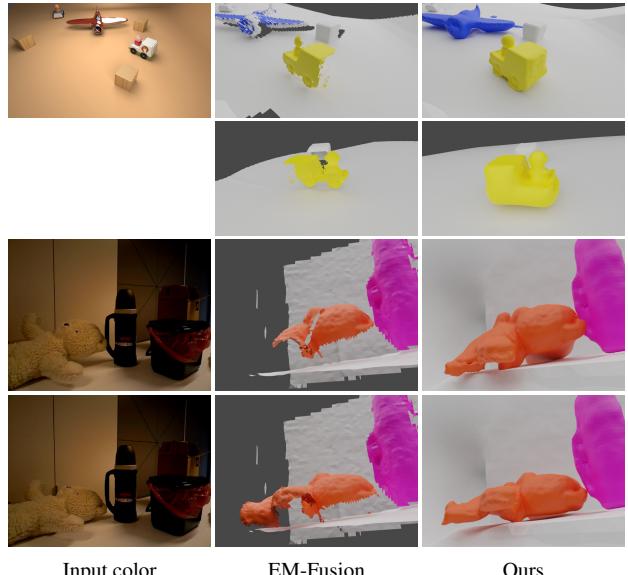


Figure 1: We propose a novel energy minimization approach to 3D reconstruction in dynamic scenes. Our approach uses a dynamic object-level SLAM method as frontend (such as EM-Fusion [19]). Compared to the TSDF object maps of the frontend, our approach can leverage intersection constraints between objects and visibility constraints to complete object shapes in a physically plausible way. The top two rows show the same timestep in the scene *car4* from different viewpoints. The bottom rows show different timesteps from the *place-items* scene. See also the supplemental video for further visualization.

The oriented point measurements are provided by a dynamic object-level SLAM frontend. It detects, segments, tracks and 3D reconstructs the objects in local volumetric SDF representations. Our energy minimization backend optimizes the object maps and incorporates intersection and hull constraints to complete the object shapes.

We demonstrate our approach on real and synthetic dynamic scene datasets. We also assess our shape completion

quantitatively using the ground-truth 3D models on the synthetic scenes.

In summary, our contributions are

- We include intersection constraints for physically plausible shape completion in dynamic scenes. This allows for inferring hidden surfaces of objects which can be useful for robotics or VR/AR applications. To the best of our knowledge, we are the first to include such constraints for object-level 3D reconstruction in dynamic scenes.
- We demonstrate efficient optimization by initializing new object volumes coarse-to-fine and optimizing the volumes incrementally when new measurements are available.
- We assess our approach qualitatively and quantitatively on dynamic SLAM datasets. Our evaluation can serve as a baseline for future research in this area.

## 2. Related work

### 2.1. Dense surface reconstruction

Several approaches reconstruct 3D surfaces densely from point clouds [2, 11, 16, 21], depth images [4] or RGB images from multiple view points [3]. The methods typically consider the data as belonging to a single surface and do not reason about the surfaces of individual objects.

**Dense reconstruction from point clouds.** Moving least squares surface reconstruction [1] fits a point set representing the surface to oriented point samples. Local parametrizations such as planar surface patches are assumed and the points are fit using moving least squares. Smooth signed distance surface reconstruction [2] recovers a volumetric implicit surface representation from the oriented points using a variational energy minimization formulation. The energy includes data terms favoring the points lying on the zero level set of the optimized implicit function and aligning its gradients with the surface normals. A regularization term keeps the Hessian, *i.e.* the second-order derivatives, of the implicit function small. While we also use this regularization term, we use a data term that measures the distance of a voxel towards a point along the normal. In Poisson surface reconstruction [11], a smoothed indicator function is estimated whose gradient follows the surface normal field approximated with the oriented points. The implicit function is recovered using a least squares approximation of the Poisson equation. Implicit moving least squares formulations recover an implicit surface function at a set of points or voxel centers using a quadratic penalty on the deviation from local point samples along their normals. Shen *et al.* [17] estimate a fit of linear function coefficients to point samples. Our formulation extends the Hessian-IMLS formulation of [16] which combines a quadratic data

penalty term with the L2-Hessian regularizer to estimate the signed distance function on a grid. The approach by Umehnofer *et al.* [21] uses L1 norms on the data terms of SSD and a total variation regularizer on the gradient of the vector field which affords a primal-dual optimization scheme. We favor the L2-Hessian regularization of the implicit function over total variation regularization of its vector field, since we are interested in recovering smooth implicit surface functions without discontinuities in order to complete object shapes watertight.

#### Dense reconstruction from depth and RGB images.

Several approaches fuse depth images which can be recovered using multi-view stereo or active sensing principles. A popular approach for incremental fusion of depth maps into volumetric signed distance functions is the seminal approach by Curless and Levoy [4]. Cremers and Kolev [3] optimize the signed distance function directly from intensity images using multi-view stereo and silhouette constraints. These dense reconstruction methods are restricted to static scenes and do not incorporate intersection constraints between objects like our method.

### 2.2. Dense object-level 3D reconstruction in dynamic environments

Several methods have been proposed that track and map moving objects with RGB-D cameras [14, 15, 23, 7, 19]. None of the methods, however, considers mutual constraints between the objects for inferring hidden surfaces at their intersections. Some recent works also consider interactions between objects, for instance, between human bodies and the static environment [8], or hands and objects [9]. However, these methods require learned parametric deformation models of bodies and hands. They furthermore rely on deep learning for pose and shape regression from images.

### 2.3. Shape completion

Various approaches have been proposed that learn to complete volumetric reconstructions obtained with depth cameras [6, 18, 5, 24]. Firman *et al.* [6] train random forests to map depth image patches to local 3D voxel representations. Song *et al.* [18] complete voxel representation of single depth images. They train a 3D convolutional neural network end-to-end on a synthetic 3D scene dataset. Scan-Complete [5] applies 3D CNNs in a hierarchical fashion to complete and semantically label voxel grids of point clouds fusing multiple views. Yang *et al.* [24] propose a GAN-style approach to predict completed shapes from depth images. The recently proposed X-Section [13] predicts the end-point of an object along a ray which can be used with volumetric SDF fusion [4] to obtain completed shapes.

Our energy minimization approach includes physically plausible constraints which do not need to be learned from data. It thus provides an orthogonal approach that might

lead to promising combinations with these learning-based approaches in future research.

### 3. Method

Our goal is to use depth measurements that are associated to object-level 3D maps in order to generate a globally consistent 3D model for each object. Our method reasons about occluded 3D surfaces by integrating intersection constraints between the objects. The key idea behind this is that dynamic objects reveal some information about their maximum extend in the viewing direction if they pass in front of another known surface. We use this information to infer where an objects ends in directions that have not been directly observed.

Our approach uses a frontend which performs dynamic object-level RGB-D SLAM. The frontend provides a segmentation of the scene into dynamic objects. It estimates the object trajectories and 3D maps which are fused from the depth measurements on the objects. Our main novelty in this work is an energy minimization backend which consolidates the dense 3D reconstructions of the moving objects with physical plausibility constraints between the objects. To this end, it takes intersection constraints between the objects into account. We also include hull constraints which limit the object surface with the observed free space.

#### 3.1. Frontend

The dynamic SLAM frontend associates depth measurements to object-level maps. RGB-D SLAM methods can be used for this purpose that track and reconstruct moving objects such as [14, 15, 23, 7, 19]. We choose to build upon our dynamic SLAM approach EM-Fusion [19] and extend it to generate globally consistent object models, including priors about object intersections. In EM-Fusion, we represent individual objects as truncated signed distance fields (TSDFs). We find the signed distance field (SDF) representation especially useful for the goal of including intersection constraints since it contains distance information to the closest surfaces. Additionally, it gives information whether a 3D point is inside or outside an object by negative or positive values, respectively.

In EM-Fusion, objects are detected by Mask R-CNN [10] and initialized as per-object TSDF volumes. For correct associations of depth pixels to object models, in each frame a geometric association likelihood  $a^o$  is estimated and used as a per-pixel weight for tracking and mapping the different models using the input depth maps. New depth measurements in EM-Fusion are integrated by weighted average fusion [4] with a cap on the maximum weight as in [12]. This approach has been proven to yield a maximum likelihood estimate of the surface [4]. However, this method cannot map thin structures and edges of objects well if they are viewed from different sides. Previously observed sur-

faces might be overridden with later measurements from the other side of the object, if they lie within the truncation distance. To alleviate these problems, we store depth frames as keyframes in equidistant timesteps and compute oriented pointclouds (consisting of points  $\mathbf{p}_i$  and their corresponding normals  $\mathbf{n}_i$ ) from them, given the poses estimated by EM-Fusion. Then we perform a global optimization of the signed distance field from this pointcloud to generate a globally consistent model. An overview of the structure of EM-Fusion and what data we use in our optimization is given in Fig. 2. We will now explain the details of the optimization framework we employ.

#### 3.2. Global SDF optimization

**Energy formulation.** Several approaches have been proposed for global optimization of signed distance fields (SDFs) from oriented pointclouds, *e.g.*, [2, 11]. Schroers *et al.* [16] proposed a taxonomy and generalization of these approaches. In our work, we adopt the Hessian-IMLS energy formulation by Schroers *et al.* [16]. The overall energy that we optimize is the sum of four parts which we will explain in the following:

$$E(u) = E_{\text{data}}(u) + E_{\text{reg}}(u) + E_{\text{hull}}(u) + E_{\text{inter}}(u). \quad (1)$$

Given measurement points  $\mathbf{p}_i$  and their corresponding normals  $\mathbf{n}_i$ , Hessian-IMLS defines the energy  $E$  over possible SDFs  $u : \Omega \rightarrow \mathbb{R}$  as

$$\begin{aligned} E_{\text{Hessian-IMLS}}(u) &= E_{\text{data}}(u) + E_{\text{reg}}(u) \\ E_{\text{data}}(u) &= \int_{\Omega} \sum_{i=1}^N w_i(u(\mathbf{x}) - f_i(\mathbf{x}))^2 d\mathbf{x} \quad (2) \\ E_{\text{reg}}(u) &= \alpha \int_{\Omega} \|\mathbf{H}u(\mathbf{x})\|_F^2 d\mathbf{x} \end{aligned}$$

with  $\Omega \subset \mathbb{R}^3$ . Here,  $f_i(\mathbf{x}) = \langle \mathbf{x} - \mathbf{p}_i, \mathbf{n}_i \rangle$  denotes the signed distance measurement for point  $\mathbf{p}_i$  along the surface normal  $\mathbf{n}_i$ . The weight  $w_i(\mathbf{x}) = w_{\sigma}(\|\mathbf{x} - \mathbf{p}_i\|) \cdot a^o(\mathbf{p}_i)$  is computed as the product of a Gaussian weight  $w_{\sigma}(s) = \exp((s/\sigma)^2)$  reducing the influence of points for voxels further from the surface as in [16] and the association likelihood  $a^o(\mathbf{p}_i)$  of point  $\mathbf{p}_i$  to object  $o$  from EM-Fusion. We set  $\sigma$  equal to the voxel size in our experiments and cap the weights when the distance to the measurement is larger than  $3\sigma$ . The L2-regularizer on the Hessian of  $u$ ,  $\|\mathbf{H}u\|_F^2$  follows the intuition that the distance to the closest surface changes approximately linearly with a change in position.

**Hull constraint.** Schroers *et al.* also propose a hull constraint which can be included into Hessian-IMLS,

$$E_{\text{hull}}(u) = \beta_{\text{hull}} \int_{\Omega \setminus \mathcal{H}} \max\{0, d_{\text{hull}}(\mathcal{H}, \mathbf{x}) - u(\mathbf{x})\}^2 d\mathbf{x}, \quad (3)$$

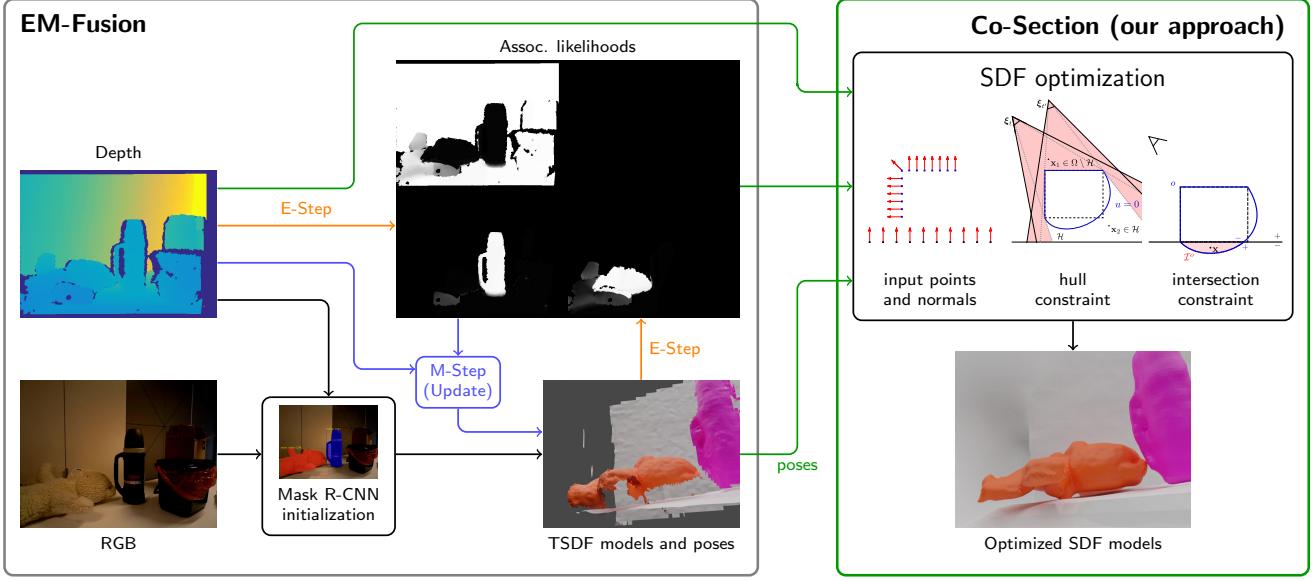


Figure 2: Our approach uses input depth maps and the estimated object and camera poses as well as the pixel-wise object association likelihoods from EM-Fusion [19] in a global optimization framework to retrieve optimized SDF models that take physical plausibility constraints into account. EM-Fusion uses Mask R-CNN [10] segmentations to initialize objects and subsequently tracks and maps them in an EM-like framework by estimating geometric association likelihoods from existing models and input depth maps.

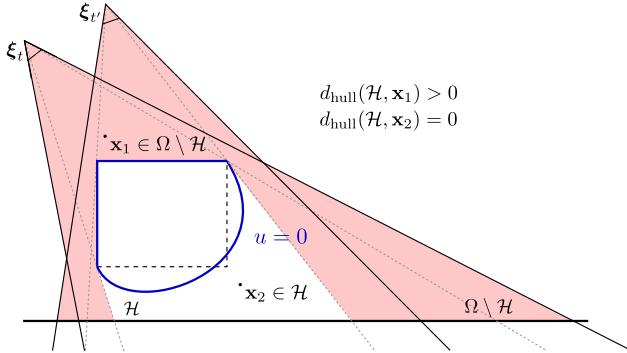


Figure 3: The hull constraint effectively limits the reconstructed surface within the observed free-space.

where  $d_{\text{hull}}(\mathcal{H}, \mathbf{x})$  denotes the distance of point  $\mathbf{x}$  to the hull  $\mathcal{H} \subset \mathbb{R}^3$  in which the reconstruction lies. Essentially, this constrains the reconstructed implicit surface to a reasonable area and proves beneficial in their experiments [16]. Since the data we collect is not just the raw oriented point cloud, but depth measurements over time, we implement the hull  $\mathcal{H}$  to consist of all unseen parts of the scene (hidden behind seen surfaces or outside the field of view). Furthermore, to avoid the potentially expensive exact computation of  $d_{\text{hull}}(\mathcal{H}, \mathbf{x})$ , we use the voxel size as a lower bound for all points  $\mathbf{x} \notin \mathcal{H}$ . This effectively limits the reconstructed

surface with the observed free space. See Fig. 3 for an illustration of the hull constraint.

**Intersection constraint.** We propose to implement an intersection constraint for multiple objects with a similar form like the hull constraint in equation (3). For each object  $o$  in the set of objects  $\mathcal{O}$ , we define the intersection distance  $d_{\text{inter}}$  based on the observations of multiple (moving) objects in several timesteps  $t$ :

$$d_{\text{inter}}^o(\mathbf{x}) = \max \left\{ \max_{t \in T, p \in \mathcal{O}} \{-u^p(\mathbf{x}_t)\}, 0 \right\}. \quad (4)$$

Here,  $\mathbf{x}_t = \mathbf{T}(\xi_t^p)^{-1} \mathbf{T}(\xi_t^o) \mathbf{x}$  denotes the point  $\mathbf{x}$  transformed from the coordinate system of object  $o$  to object  $p$  at timestep  $t$  using the poses of  $o$  and  $p$  at that timestep,  $\xi_t^o$  and  $\xi_t^p$ , respectively. Intuitively, this distance measures the maximum penetration of a point  $\mathbf{x}$  of object  $o$  in object  $p$  over all time steps.

Using this distance, we define the energy term

$$E_{\text{inter}}(u^o) = \beta_{\text{inter}} \int_{\mathcal{I}^o} \max\{0, d_{\text{inter}}^o - u^o\}^2 d\mathbf{x}, \quad (5)$$

where  $\mathcal{I}^o = \{\mathbf{x} \in \Omega \mid d_{\text{inter}}^o(\mathbf{x}) > 0\}$ . This term favors SDFs that do not interpenetrate each others. Fig. 4 illustrates the intersection constraint in several situations.

**Optimization.** Since all parts of the energy are convex and differentiable in  $u$ , we follow Schroers *et al.* [16] and implement the optimization with the Fast Jacobi algorithm proposed by Weickert *et al.* [22].

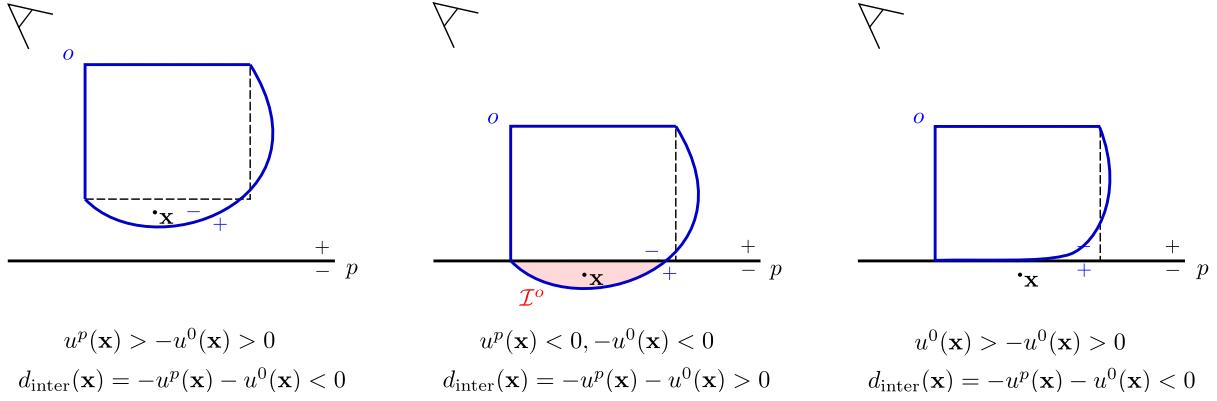


Figure 4: Our intersection constraint is based on a measure of interpenetration distance  $d_{\text{inter}}$ . It penalizes signed distance functions that interpenetrate each other. Left: interpenetration distance for two objects that do not intersect each others. Middle: case of two intersecting objects. Right: Distance after resolved interpenetration.

Since in our case the data measurements are sparse compared to the overall volume size, optimization of the background volume on the finest resolution of  $256^3$  only converges slowly for the first run. We thus propose to initialize each volume with a coarse-to-fine optimization by optimizing only  $E_{\text{data}}$  from equation (2) first on a resolution of  $32^3$  and subsequently upsampling the result by splitting each voxel into 8 and using this as initialization for the next level. Even without interpolation, this strategy yielded reasonable fast and accurate results. For the finest level, we optimize equation (1) and include the hull and intersection constraints. The coarse-to-fine scheme is only applied for the first optimization of each volume. In our experiments, the previous optimization result can adapt reasonably fast to new measurements in later optimization runs. Furthermore, we found the application of a coarse-to-fine scheme in later time steps to actually slow down the optimization in some cases. This is caused by removing fine details at lower resolutions which were already recovered in the previous optimization run.

**Wrong data and inaccurate detections.** The detection of dynamic objects in [19] relies on Mask R-CNN [10]. The detection is carried out at low framerate. This might lead to dynamic objects being missed when they move into view. Consequently, point associations with the background or other object models are incorporated which are difficult to remove later. Since EM-Fusion is mainly designed to track rigid objects, it is fair to assume that space that was seen unoccupied before in a model will not be occupied by the same object in the future. We thus do not allow association of points in this space with the respective model and set  $f_i(\mathbf{x}) = w_i(\mathbf{x}) = 0$  for  $\mathbf{x} \in \Omega \setminus \mathcal{H}$ . Note that this way, the hull  $\mathcal{H}$  will only get smaller over time.

Furthermore, EM-Fusion yields balanced association likelihoods for the background model and objects that re-

main static since the first view. Including these points in both the object and the background model during our optimization would yield conflicting dataterms for the intersection constraint. To avoid this, we use the volumetric foreground probability estimated by EM-Fusion and remove all point measurements in the foreground region of an object model from all other models. While this strategy helps with handling static objects, it might cause dents when models touch each others since it might also remove correct point measurements from the other model. This strategy might not be needed in scenes where no static objects are detected by EM-Fusion.

**Ordering the optimization of different volumes.** The intersection constraint as formulated in equation (4) introduces a chicken-and-egg-problem. The optimization for object  $o$  depends on the optimization results for all other models while those results depend on the optimization result of  $o$ . To resolve this, we observe that planar surfaces (as they are common for e.g. walls and floors in manmade environments) can be completed reasonably well using the Hessian norm prior in equation (2). Thus to compute the intersection constraint in Eq. (4), we use the optimized SDF of the background model and approximate  $u^p$  in the case of object models using the point measurements, i.e.  $u^p(\mathbf{x}) \approx \frac{1}{N} \sum_{i=1}^N w_i(\mathbf{x}) f_i(\mathbf{x})$ .

## 4. Experiments

We evaluate our approach in qualitative and quantitative experiments on the dataset provided with Co-Fusion [14]. We use 2 real and 1 synthetic scene in which up to 3 objects move independently. Besides qualitative results, we also provide quantitative results on shape completion which we obtain in the synthetic scene using the ground-truth meshes of the objects. In an ablation study, we analyze the effec-

tiveness of the individual components of our energy minimization approach.

We chose the parameters for weighting the different parts of the energy empirically as  $\alpha = 0.005$ ,  $\beta_{\text{hull}} = \beta_{\text{inter}} = 0.001$ , and choose a cycle length of 20 for the Fast Jacobi optimizer [22]. We used a single set of parameters throughout our experiments.

## 4.1. Qualitative Results

We provide qualitative results of our approach in Fig. 5. The first three examples show different viewpoints of the same time step, while the last one (placement of teddy) illustrates how the shape changes over time. The TSDF reconstructions estimated by EM-Fusion can only take the visible part of the object into account. When using our global optimization approach without hull and intersection constraints, the object surface is continued as smooth as possible towards the borders of the object map. The hull constraints wraps the surfaces such that they are limited by the observed free space. However, the surface can freely intersect with the background surface under the object. When including the intersection constraint, this penetration is avoided. However, without the hull constraint, the gradient introduced by the intersection constraint might be continued in regions outside the actual objects, causing unwanted zero-crossings (see column 4 in Fig. 5). In combination with the hull constraint, these unwanted zero-crossings are removed by penalizing negative SDF values in observed free space. Thus, the object shape is closed and approximates the actual object shape well.

More qualitative examples including the development of the optimized surfaces over time can be found in the supplemental video.

## 4.2. Quantitative Results

**Evaluation metric.** We evaluate shape reconstruction using the measures and tools suggested in [20]<sup>1</sup>. The method samples 10,000 points on the mesh reconstructed from the SDFs and the ground-truth mesh uniformly. It then computes the average distance of each sample point in the reconstructed mesh to the ground-truth mesh (point-to-triangle distance) as *accuracy*, and the average distance of each sample from the ground-truth mesh to the reconstructed mesh as *completeness*. Unfortunately, for this evaluation the alignment of the ground-truth mesh is not provided by the dataset. Thus, we manually align scale and pose of the ground-truth meshes with our output meshes by selecting point correspondences on the meshes. Note that, we only need to align each object once for each optimized map (using the full approach), since all reconstructions are based on the EM-Fusion result and share the same pose estimate.

---

<sup>1</sup><https://github.com/davidstutz/mesh-evaluation>

**Results.** Table 1 lists accuracy and completeness for the objects on the synthetic car4 sequence of the Co-Fusion dataset. Our full optimization approach strongly improves the completeness of the shape towards the TSDF map. For two of three objects, accuracy is reduced towards the TSDF map. However, this is expected since our method inpaints new object parts which are not modelled by the TSDF. Still, the accuracy of our full approach is better than any variant of the optimization method. We observe that while the hull constraint tends to improve accuracy, the intersection constraint tends to improve completeness. We conclude that both the hull and intersection constraints are important to achieve the performance of the full model. In Fig. 6 we show accuracy and completeness per mesh vertex (point-to-triangle distance).

Our method relies on accurate pose estimates by the underlying dynamic SLAM method. In case of inaccurate pose estimates, hull and intersection volumes might become inaccurate as well, leading to inaccuracies in the surface reconstruction. While we have not observed this problem in our experiments, a stronger weighting of the dataterm  $E_{\text{data}}$  (2) could alleviate this problem to some extent.

**Computation times.** The intersection constraint can be computed for several voxels in parallel on a GPU. In our experiments this computation takes under 10ms on average per volume. Similarly, the data measurements and the hull constraint can be entered fast on parallel hardware. The optimization itself is more time consuming. For the object volumes in our experiments, our implementation of our variant of the Hessian-IMLS [16] achieves runtimes of several seconds (avg. 0.98s, peak 19.8s). The peak runtime typically occurs when a mostly empty volume needs to be filled from sparse measurements after initialization. We consider further parallelization and improvement of the runtime efficiency of our approach as future work. Schroers *et al.* [16] report runtimes between 1s and 4s for volumes with resolutions up to  $400^3$  (our resolution  $64^3/256^3$  for objects/background, respectively).

## 5. Conclusion

In this paper we propose a novel energy minimization approach for object shape completion in dynamic scenes. We incorporate hull and intersection constraints between objects into a formulation which optimizes for the implicit surface in a volumetric representation. The data terms of our method are obtained with a dynamic object-level SLAM frontend which detects, segments, tracks and maps the objects in local TSDF volumes. We demonstrate in our experiments that our formulation can achieve object shape completion which is physically plausible. We analyze the contributions of the constraints for accuracy and completeness of object shape reconstruction. Our approach improves completeness over the TSDF reconstruction and can achieve

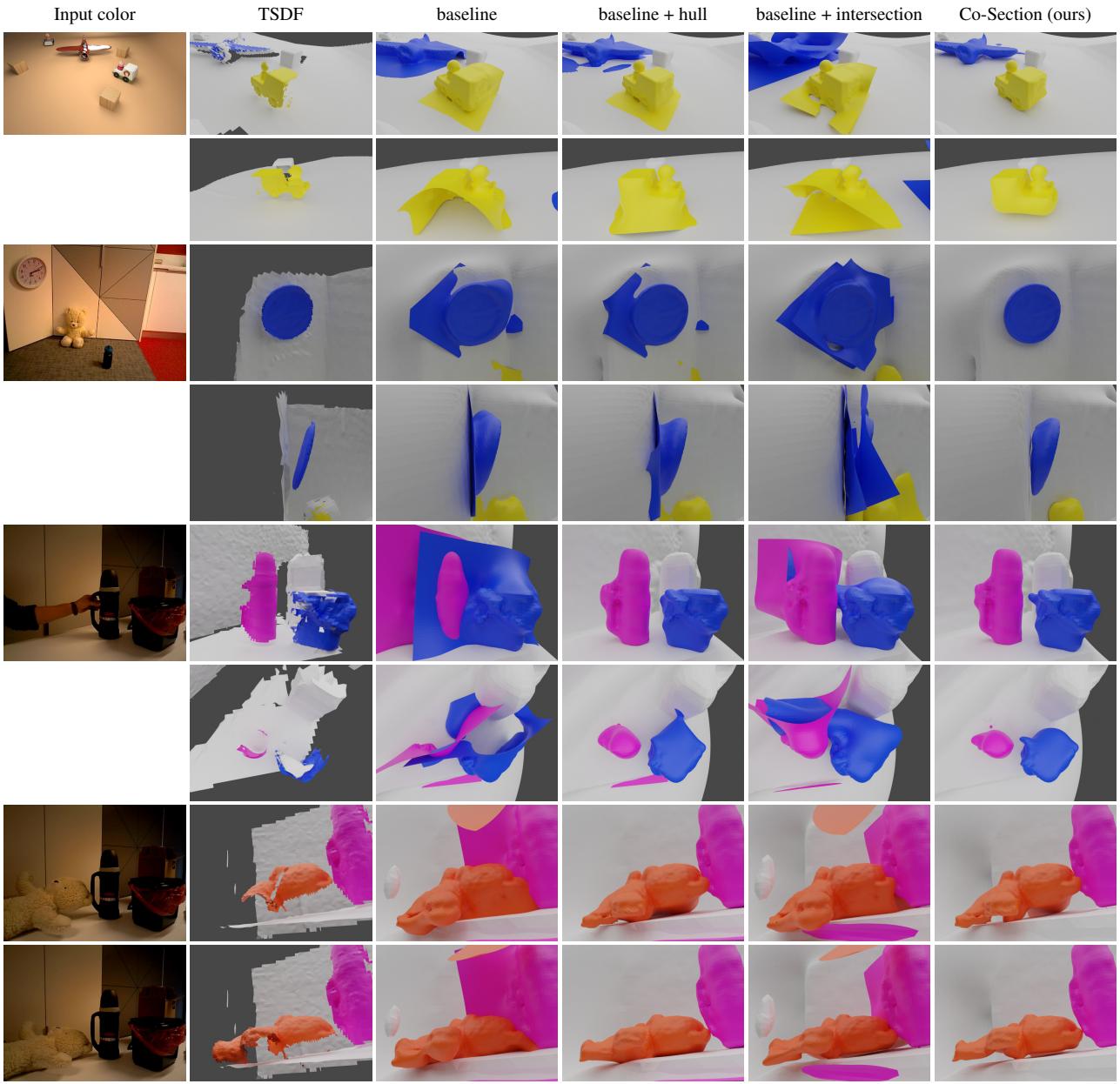


Figure 5: Qualitative object shape reconstruction results on Co-Fusion sequences. From top to bottom: truck (car4), clock (sliding-clock), bottle and dustbin (place-items), and teddy (place-items). The TSDF shape only includes the observed part of the objects. While the hull constraint limits the surface in the observed free space, the intersection constraint closes the object shape at intersections with other shapes. Our full approach recovers closed surfaces which approximate the actual object shapes. Note that we do not show “person” objects such as the arm in the bottle and dustbin example.

high accuracy even if parts of the object are unobserved.

In future work, we would like to investigate the incorporation of further regularization constraints to achieve improved scene reconstruction. Currently, our method is not real-time capable. Optimization of a volume takes from under a second to several seconds. This can still be interest-

ing for backend optimization in a parallel thread with the frontend. Devising methods for increasing the run-time efficiency is an interesting direction for future research.

**Acknowledgements.** We acknowledge support from Cyber Valley, the Max Planck Society, and the German Fed-

Method	Truck		Car		Airplane	
	Acc	Comp	Acc	Comp	Acc	Comp
TSDF	0.00949544	0.0375519	<b>0.00204528</b>	0.0277146	<b>0.0364824</b>	0.0281357
baseline	0.0333811	0.0327722	0.0154461	0.0196067	0.156872	0.0229247
baseline + hull	0.0234413	0.023931	0.0117728	0.0122979	0.150925	0.0174758
baseline + intersection	0.0325838	0.0116423	0.0163322	0.0109024	0.173195	0.0202907
Co-Section (ours)	<b>0.00878659</b>	<b>0.0109651</b>	0.00521935	<b>0.0102169</b>	0.0871882	<b>0.016172</b>

Table 1: Accuracy and completeness (lower is better) on the Co-Fusion car4 sequence for different variants of our method. Best in bold. Our full approach is clearly superior in completeness and also performs well in accuracy despite the fact that it “guesses” parts of the object surface from the constraints.

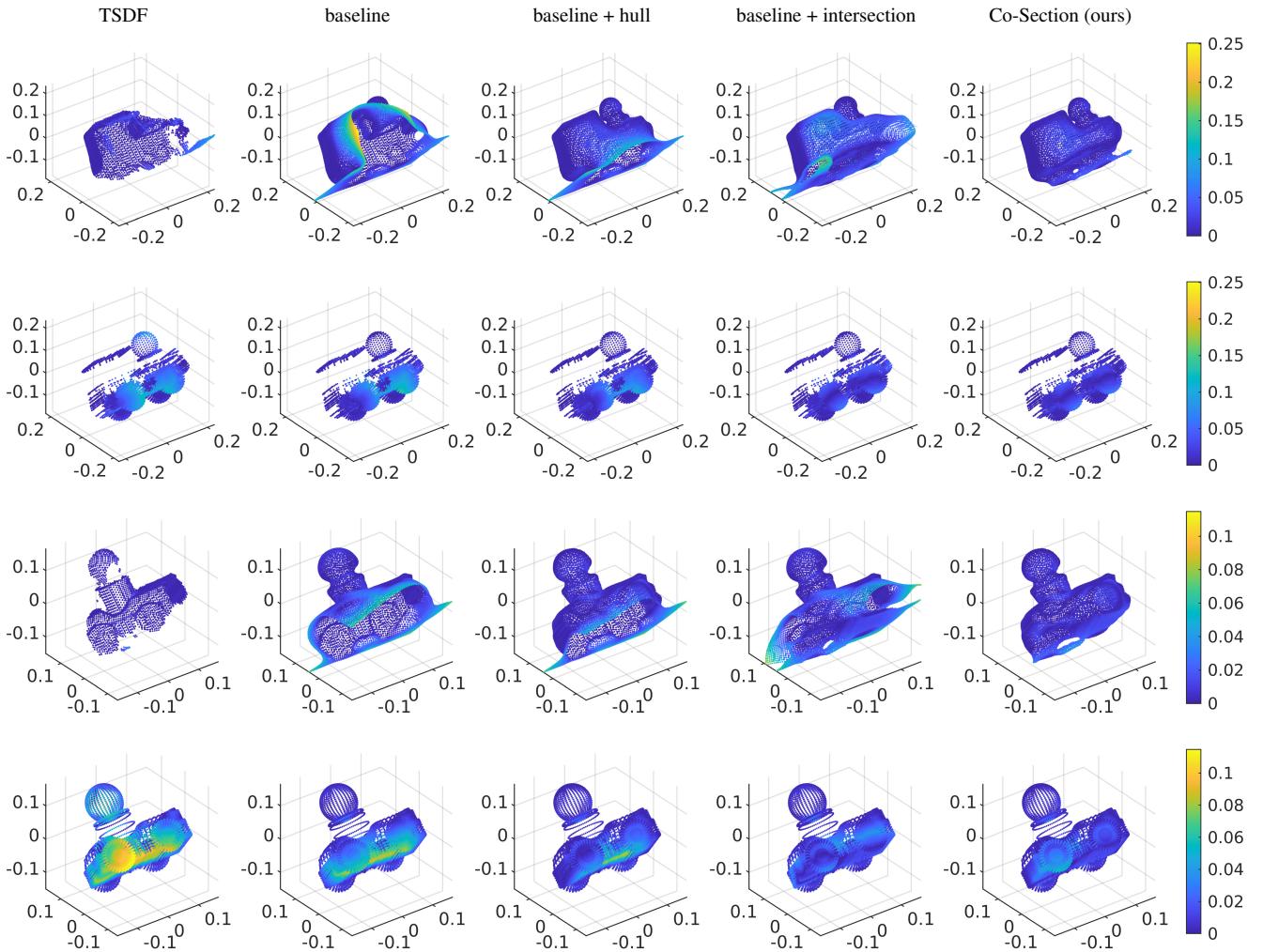


Figure 6: Accuracy (top row per object) and completeness (bottom row per object) of object shape reconstruction results on the Co-Fusion car4 sequence. Color indicates distance of mesh vertices to the other mesh. From top to bottom: truck, car. Our full approach is clearly superior to all other variants in accuracy as well as completeness.

eral Ministry of Education and Research (BMBF) through the Tuebingen AI Center (FKZ: 01IS18039B). The authors thank the International Max Planck Research School

for Intelligent Systems (IMPRS-IS) for supporting Michael Strecke.

## References

- [1] M. Alexa, J. Behr, D. Cohen-Or, S. Fleishman, D. Levin, and C.T. Silva. Computing and rendering point set surfaces. *IEEE Transactions on Visualization and Computer Graphics*, 9(1):3–15, jan 2003. [2](#)
- [2] Fatih Calakli and Gabriel Taubin. SSD: Smooth signed distance surface reconstruction. *Computer Graphics Forum*, 30(7):1993–2002, sep 2011. [2](#), [3](#)
- [3] Daniel Cremers and Kalin Kolev. Multiview stereo and silhouette consistency via convex functionals over convex domains. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1161–1174, jun 2011. [2](#)
- [4] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, SIGGRAPH ’96, pages 303–312, New York, NY, USA, 1996. ACM Press. [2](#), [3](#)
- [5] Angela Dai, Daniel Ritchie, Martin Bokeloh, Scott Reed, Jürgen Sturm, and Matthias Nießner. ScanComplete: Large-scale scene completion and semantic segmentation for 3D scans. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2018. [1](#), [2](#)
- [6] Michael Firman, Oisin Mac Aodha, Simon Julier, and Gabriel J. Brostow. Structured prediction of unobserved voxels from a single depth image. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2016. [1](#), [2](#)
- [7] Ryo Hachiuma, Christian Pirchheim, Dieter Schmalstieg, and Hideo Saito. DetectFusion: Detecting and segmenting both known and unknown dynamic objects in real-time SLAM. In *Proceedings British Machine Vision Conference (BMVC)*, 2019. [2](#), [3](#)
- [8] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. [2](#)
- [9] Yana Hasson, Gülcin Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2019. [2](#)
- [10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988. IEEE, oct 2017. [3](#), [4](#), [5](#)
- [11] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*, SGP ’06, pages 61–70, Aire-la-Ville, Switzerland, Switzerland, 2006. Eurographics Association. [2](#), [3](#)
- [12] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 127–136. IEEE, oct 2011. [3](#)
- [13] Andrea Nicastro, Ronald Clark, and Stefan Leutenegger. X-Section: Cross-section prediction for enhanced RGB-D fusion. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. [2](#)
- [14] Martin Rünz and Lourdes Agapito. Co-Fusion: Real-time segmentation, tracking and fusion of multiple objects. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4471–4478. IEEE, may 2017. [2](#), [3](#), [5](#)
- [15] Martin Rünz, Maud Buffier, and Lourdes Agapito. MaskFusion: Real-time recognition, tracking and reconstruction of multiple moving objects. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 10–20. IEEE, oct 2018. [2](#), [3](#)
- [16] Christopher Schroers, Simon Setzer, and Joachim Weickert. A variational taxonomy for surface reconstruction from oriented points. *Computer Graphics Forum*, 33(5):195–204, aug 2014. [2](#), [3](#), [4](#), [6](#)
- [17] Chen Shen, James F. O’Brien, and Jonathan R. Shewchuk. Interpolating and approximating implicit surfaces from polygon soup. *ACM Transactions on Graphics*, 23(3):896, aug 2004. [2](#)
- [18] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jul 2017. [1](#), [2](#)
- [19] Michael Strecke and Jörg Stückler. EM-Fusion: Dynamic object-level SLAM with probabilistic data association. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. [1](#), [2](#), [3](#), [4](#), [5](#)
- [20] David Stutz and Andreas Geiger. Learning 3D shape completion from laser scan data with weak supervision. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2018. [6](#)
- [21] Benjamin Ummenhofer and Thomas Brox. Global, dense multiscale reconstruction for a billion points. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, dec 2015. [2](#)
- [22] Joachim Weickert, Sven Grewenig, Christopher Schroers, and Andrés Bruhn. Cyclic schemes for PDE-based image analysis. *International Journal of Computer Vision*, 118(3):275–299, dec 2015. [4](#), [6](#)
- [23] Binbin Xu, Wenbin Li, Dimos Tsoumanikas, Michael Bloesch, Andrew Davison, and Stefan Leutenegger. MID-Fusion: Octree-based object-level multi-instance dynamic SLAM. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, may 2019. [2](#), [3](#)
- [24] Bo Yang, Stefano Rosa, Andrew Markham, Niki Trigoni, and Hongkai Wen. Dense 3D object reconstruction from a single depth view. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):2820–2834, dec 2019. [1](#), [2](#)