

ReFusion: 3D Reconstruction in Dynamic Environments for RGB-D Cameras Exploiting Residuals

Emanuele Palazzolo

Jens Behley

Philipp Lottes

Philippe Giguère

Cyrill Stachniss

Abstract— Mapping and localization are essential capabilities of robotic systems. Although the majority of mapping systems focus on static environments, the deployment in real-world situations requires them to handle dynamic objects. In this paper, we propose an approach for an RGB-D sensor that is able to consistently map scenes containing multiple dynamic elements. For localization and mapping, we employ an efficient direct tracking on the truncated signed distance function (TSDF) and leverage color information encoded in the TSDF to estimate the pose of the sensor. The TSDF is efficiently represented using voxel hashing, with most computations parallelized on a GPU. For detecting dynamics, we exploit the residuals obtained after an initial registration, together with the explicit modeling of free space in the model. We evaluate our approach on existing datasets, and provide a new dataset showing highly dynamic scenes. These experiments show that our approach often surpass other state-of-the-art dense SLAM methods. We make available our dataset with the ground truth for both the trajectory of the RGB-D sensor obtained by a motion capture system and the model of the static environment using a high-precision terrestrial laser scanner. Finally, we release our approach as open source code.

I. INTRODUCTION

Mapping and localization are essential capabilities of robotic systems operating in real-world environments. Simultaneous localization and mapping, or SLAM, is usually solved in an alternating fashion, where one determines the pose w.r.t. the map built so far and then use the estimated pose to update the map. SLAM is especially challenging in dynamic environments, since a robot needs to build a consistent map. This entails estimating simultaneously which parts of the environment are static or moving. In particular, moving objects may cause wrong correspondences, deteriorate the ability to estimate correct poses and hence corrupt the map.

In this paper, we propose *ReFusion*: a novel approach for dense indoor mapping that is robust to dynamic elements. Our approach is completely geometric, and does not rely on an explicit semantic interpretation of the scene. More specifically, we do not employ a deep neural network to detect specific dynamic classes, in contrast to other recent approaches [13], [1]. In contrast to other recent purely geometric approaches [12], [14], we do not represent the model using surfels, but in the form of a truncated sign distance function (TSDF). This allows our technique to directly generate a mesh of the environment, which can be useful in

Emanuele Palazzolo, Jens Behley, Philipp Lottes and Cyrill Stachniss are with the University of Bonn, Germany. Philippe Giguère is with the Laval University, Québec, Canada.

This work has partly been supported by the DFG under the grant number FOR 1505: Mapping on Demand, under the grant number BE 5996/1-1, and under Germany’s Excellence Strategy, EXC-2070 - 390732324 (PhenoRob).

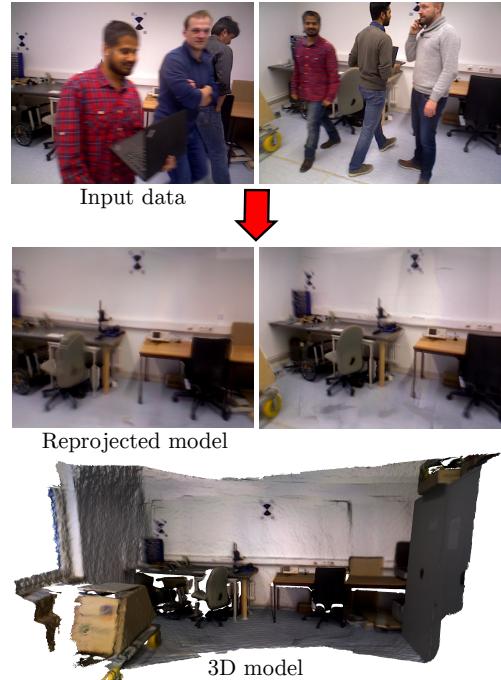


Fig. 1: Result of our approach. Top: RGB frames from our dataset containing dynamics. Center: Reconstructed model without dynamics reprojected onto the two frames from above. Bottom: Final mesh without the dynamics.

application such as virtual and augmented reality. Moreover, the TSDF representation can be useful for planning, since it provides, by definition, the distance to the closest obstacle.

The main contribution of this paper is a novel and efficient SLAM algorithm, based on a TSDF representation, that is robust to dynamics via pure geometric filtering. It is inspired by the work of Canelhas *et al.* [3] combined with voxel hashing [10]. We propose to detect dynamics by exploiting the residuals obtained from the registration, in combination with the explicit representation of free space in the environment. This allows our approach to be class agnostic, i.e., it does not rely on a detector trained on specific categories of dynamic objects. Moreover, the dynamic objects are not explicitly tracked, therefore our approach is not limited by the number of moving objects or by their speed. Fig. 1 shows two example RGB frames from a dynamic scene and the resulting model built by our approach.

We evaluated ReFusion on the TUM RGB-D dataset [17], as well as on our own dataset, showing the versatility and robustness of our approach, reaching in several scenes equal or better performance than other dense SLAM approaches. In addition, we publicly release our dataset, containing 24

highly-dynamic scenes recorded with an RGB-D sensor, together with ground truth trajectories obtained using a motion capture system. Furthermore, we provide a ground truth 3D model of the static parts of the environment in the form of a high resolution point cloud acquired with a terrestrial laser scanner. To the best of our knowledge, this is the first dataset containing dynamic scenes that also includes the ground truth model for the static part of the environment. Finally, we publicly share the open source implementation of our approach.

In sum, we make two key claims: our RGB-D mapping approach (i) is robust to dynamic elements in the environment and provides a camera tracking performance on par or better than state-of-the-art dense SLAM approaches, and (ii) provides a dense 3D model that contains only the static parts of the environment, which is more accurate than other state-of-the-art approaches when compared to the ground truth model.

II. RELATED WORK

With the advent of inexpensive RGB-D cameras, many approaches for mapping using such sensors were proposed [15], [20]. The seminal paper of Newcombe *et al.* [9] showed the prospects of TSDF-based RGB-D mapping by generating accurate, high detailed maps using only depth information. It paved the way for several improvements increasing the versatility and fidelity of RGB-D mapping. As the approach relies on a fixed voxel grid, volumes that can be mapped are limited. Subsequent approaches explore compression of this grid. For instance, Steinbrücker *et al.* [16] use an Octree instead of a voxel grid. Nießner *et al.* [10], on the other hand, propose to only allocate voxel blocks close to the mapped surface, and address them in constant time via hashing. Kähler *et al.* [6] extend the idea of voxel hashing by using a hierarchy of voxel blocks with different resolutions. To alleviate the need for raycasting for generating the model image for registration, Canelhas *et al.* [3] and Bylow *et al.* [2] propose to directly exploit the TSDF for evaluation of the residuals and computation of the jacobians within the error minimization.

Besides using a TSDF, another popular representation of the model are surfels, which are disks with a normal and a radius. Keller *et al.* [7] use surfels to represent the model of the environment. Whelan *et al.* [19] extend the approach with a deformation graph, which allows for long-range corrections of the map via loop closures.

An alternative approach was proposed by Della Corte *et al.* [5]. Their approach registers two consecutive RGB-D frames directly upon each other by minimizing the photometric error. They integrate multiple cues such as depth, color and normal information in a unified way. Other approaches exploit image sequences [?], [?].

Usually, mapping approaches assume a static environment and therefore handle moving objects mostly through outliers rejection. By discarding information that disagrees with the current measurements, one can handle implicitly dynamic objects [9]. Other approaches model the dynamic parts of

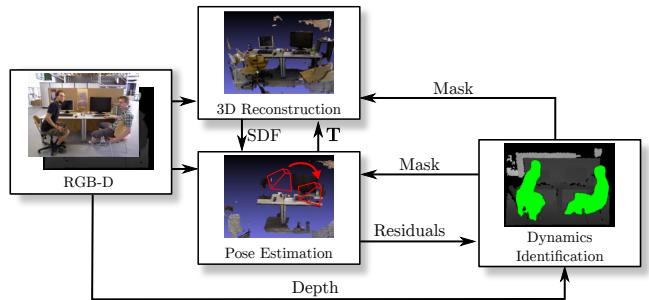


Fig. 2: Overview of our approach. Given data from the RGB-D sensor, we first perform an initial pose estimation. Then, we use the obtained residuals, together with the depth information, to identify dynamic parts of the scene. The filtered images are then used to refine the pose T w.r.t. the TSDF given by our 3D reconstruction. With the updated sensor pose, we finally integrate the measurements into our 3D reconstruction of the environment.

the environment explicitly [7], [14] and filter these before the integration into the model. Keller *et al.* [7] use outliers in point correspondences during ICP as seed for segmentation of dynamics. Corresponding model surfels inside the segments are then marked as unstable. In contrast, Rünz *et al.* [12] explicitly track moving objects given by a segmentation process using either motion or semantic cues provided by class-agnostic object proposals [11]. Scona *et al.* [14] extend ElasticFusion [19] to incorporate only clusters that correspond to the static environment. Distinguishing static and dynamic parts of the environment is achieved by jointly estimating the camera pose and whether clusters are static or dynamic. Bescos *et al.* [1] combine a geometric approach with a deep learning segmentation to enable removal of dynamics in an ORB-SLAM2 system [8]. Similarly, Rünz *et al.* [13] exploit deep learning segmentation, refined with a geometric approach, to detect objects. In addition, they reconstruct and track detected objects independently.

In contrast to the aforementioned approaches, we propose a mapping approach robust to dynamics which relies on the residuals from the registration and on the detected free space. In this way, our approach is able to detect any kind of dynamics without relying on specific classes or models, and without explicitly tracking dynamic objects.

III. OUR APPROACH

Fig. 2 illustrates the key processing steps of the proposed approach. Given the color and depth information of an RGB-D sensor, like Microsoft’s Kinect, we first perform an initial pose estimation by exploiting directly the TSDF of our model representation. By observing the residuals obtained from such registration, we detect the dynamic elements in the scene. With the filtered sensor information, where we discard regions containing dynamics, we further refine the pose of the sensor. With this refined estimated pose, we then integrate the sensor measurements, i.e., depth and color, into the model.

A. Model Representation

In our approach, we represent the model of the environment using a truncated signed distance function (TSDF) as

originally proposed by Curless and Levoy [4]. We briefly recap it here for the sake of a self-contained description. The idea is to represent the world with a 3D voxel grid in which each voxel contains a SDF value. The SDF is a function $V_{\text{SDF}}(\mathbf{x}) : \mathbb{R}^3 \rightarrow \mathbb{R}$ that returns, given a point in space, its signed distance to the nearest surface. This distance is positive if the point is in front of the surface and negative otherwise. In this way, the surface is implicitly represented by the set of points for which $V_{\text{SDF}}(\mathbf{x}) = 0$. In practice, the SDF values are truncated to a given maximum value. In addition to the TSDF, each voxel contains a weight that represents how reliable the SDF value is at that location, as well as the color obtained by projecting it onto the RGB image. The weight allows us to update each voxel using a running weighted average, leading to an improved robustness to outliers [4]. The color information enables both the texturing of the mesh and the use of intensity information during the registration.

In our implementation, we do not preallocate a 3D voxel grid of a fixed size. Instead, we allocate voxel dynamically and index them using a spatial hashing function, similarly to Nießner *et al.* [10]. Representing the world sparsely by only allocating the needed blocks of voxels enables the reconstruction of larger scenes, while eliminating the need for restricting the maximum size of the scene in advance. Note that we process every voxel in parallel on the GPU, since the voxels are assumed to be mutually independent, leading to a considerable speed-up of the computations.

B. Pose Estimation

For estimating the pose of the sensor, we use a point-to-implicit approach as proposed independently by Canelhas *et al.* [3] and Bylow *et al.* [2]. In contrast to KinectFusion [9] and similar methods, our approach does not generate synthetic views from the model. Instead, we use an alternative technique and directly align an incoming point cloud from the sensor to the SDF, since the SDF provides by definition the distance of a point to the closest surface, i.e., it is possible to directly use the SDF value as an error function. In addition to the existing point-to-implicit techniques, we exploit the color information contained in the model to improve the alignment.

Each frame of an RGB-D sensor consists of a depth image and a color image. Given a pixel $\mathbf{p} = [u \ v]^T$, we define the functions $D(\mathbf{p}) : \mathbb{R}^2 \rightarrow \mathbb{R}$ and $I(\mathbf{p}) : \mathbb{R}^2 \rightarrow \mathbb{R}$, which map a pixel to its depth and its intensity, respectively.

We denote by \mathbf{x} the 3D point resulting from the back-projection of a pixel \mathbf{p} :

$$\mathbf{x} = \begin{bmatrix} \frac{u - c_x}{f_x} D(\mathbf{p}) \\ \frac{v - c_y}{f_y} D(\mathbf{p}) \\ D(\mathbf{p}) \end{bmatrix}, \quad (1)$$

where c_x , c_y , f_x and f_y are the intrinsic parameters of the camera, assuming a pinhole camera model.

We represent a camera pose as a 3D transformation $\mathbf{T} \in \text{SE}(3)$. A small rigid-body motion can be written in minimal

form using the Lie algebra representation $\xi \in \mathfrak{se}_3$, which can be converted into the corresponding rigid transformation $\mathbf{T}' \in \text{SE}(3)$ using the exponential map $\mathbf{T}' = \exp(\hat{\xi})$, where $\hat{\xi}$ is the corresponding skew symmetric matrix of ξ .

To represent the current model, we use a voxel-based representation, where we store in each voxel the TSDF value and color information. We define the functions $V_{\text{SDF}}(\mathbf{x})$ and $V_I(\mathbf{x})$ that return respectively the SDF and the intensity from the model at position \mathbf{x} . We obtain the intensity value I from the RGB color information stored in the voxels, i.e., $I = 0.2126R + 0.7152G + 0.0722B$. We furthermore use trilinear interpolation on the SDF and intensity of neighboring voxels to alleviate discretization effects of the voxel grid.

As mentioned before, we directly exploit the TSDF to define the error function, since it directly represents the distance of a point to the nearest surface of the model. We define the error function relative to the depth as:

$$E_d(\hat{\xi}) = \sum_{i=1}^N \underbrace{\|V_{\text{SDF}}(\exp(\hat{\xi})\mathbf{T}\mathbf{x}_i)\|^2}_{r_i}, \quad (2)$$

where N is the number of pixels in the image, and $\mathbf{x}_i, i \in 1, \dots, N$, is the 3D point corresponding to the i -th pixel \mathbf{p}_i , computed using Eq. (1). The value r_i corresponds to the residual for the i -th pixel.

We additionally use the color information to improve the alignment. In contrast to most of the state-of-the-art approaches [19], [18], we do not render a synthetic view from the model. Instead, we directly operate on the intensity obtained from the color information stored in the voxels. We define the error function relative to the intensity as the photometric error between the intensity of the pixels of the current image and the intensity of the corresponding voxels in the model:

$$E_c(\hat{\xi}) = \sum_{i=1}^N \|V_I(\exp(\hat{\xi})\mathbf{T}\mathbf{x}_i) - I(\mathbf{p}_i)\|^2, \quad (3)$$

and the joint error function is given by:

$$E(\hat{\xi}) = E_d(\hat{\xi}) + w_c E_c(\hat{\xi}), \quad (4)$$

with w_c weighting the contribution of the intensity information w.r.t. the depth information.

We solve this least-squares problem using Levenberg-Marquardt on three different coarse-to-fine sub-sampling of the input images to speed-up convergence. We furthermore exploit the GPU by processing each pixel in parallel to further accelerate the minimization process.

C. Dynamics Detection

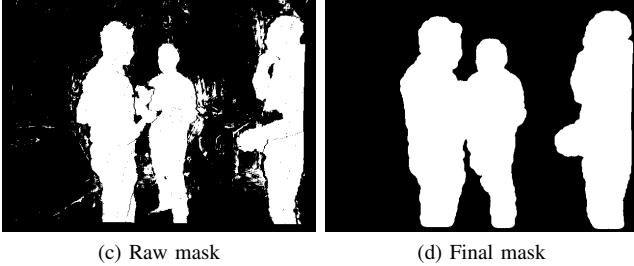
To detect dynamic parts of the environment, we first perform an initial registration of the current RGB-D frame with respect to the model, as described in Sec. III-B. After registration, we compute for each pixel \mathbf{p}_i its residual r_i w.r.t. the model as defined in Eq. (2) and illustrated in Fig. 3b. We select a threshold:

$$t = \gamma \tau^2, \quad (5)$$



(a) RGB Image

(b) Residuals



(c) Raw mask

(d) Final mask

Fig. 3: Steps of the mask creation. (a) Example RGB frame. (b) Residuals obtained from the registration. (c) Initial mask obtained from the residual. (d) Final refined mask after floodfill.

Algorithm 1: floodfill for generating \mathcal{M}_D .

Input: pixels of residual mask \mathcal{M}_R
Result: Mask \mathcal{M}_D

Let \mathcal{Q} be a queue containing all the pixels to be masked.
 Let $\mathcal{N}(\mathbf{p})$ be the set of neighbors of pixel \mathbf{p}
 Add all pixels from \mathcal{M}_R to queue \mathcal{Q}

while $\mathcal{Q} \neq \emptyset$ **do**

- Add all pixels inside \mathcal{Q} to \mathcal{M}_D
- foreach** $\mathbf{p} \in \mathcal{Q}$ **do**
- foreach** $\mathbf{n} \in \mathcal{N}(\mathbf{p})$ **do**
- if** $\|\mathcal{D}(\mathbf{p}) - \mathcal{D}(\mathbf{n})\| < \theta \cdot \mathcal{D}(\mathbf{p})$ **then**
- Add \mathbf{n} to \mathcal{Q} **if** $\mathbf{n} \notin \mathcal{M}_D$
- end**
- end**
- Remove \mathbf{p} from \mathcal{Q}
- end**

where τ is the truncation distance used in our TSDF representation and γ is a value between 0 (everything is masked) and 1 (nothing is masked). Fig. 4 shows a histogram of the residuals obtained after the initial registration of one image. The figure shows how most of the residuals concentrate below a certain value, except the ones belonging to dynamic parts of the environment. Every residual exceeding t contributes to the creation of a binary mask, see Fig. 3c. Such threshold-based segmentation is often not perfect and may fail to capture the whole dynamic object. Since we have depth information available, we use the eroded mask \mathcal{M}_R to initialize a depth-aware flood fill algorithm, summarized in Alg. 1, similar to the region growing approaches used in [7], [13]. We add neighbors to a region as long as their depth does

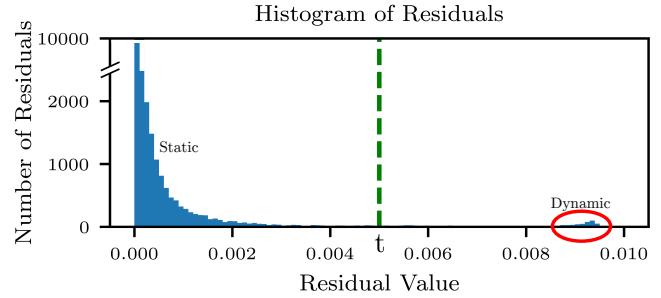


Fig. 4: Histogram of residuals obtained after the registration of an image containing dynamic elements. The red ellipse highlights the residuals resulting from the dynamic parts. By considering only pixels with residuals under the threshold t , we can identify the dynamic parts of the image.

not differ more than a threshold θ . Finally, the mask is dilated again to cover eventual border pixels left out by the floodfill. Fig. 3d shows the resulting mask after all the processing steps. We then perform a second registration without masked pixels and we integrate the RGB-D information ignoring the masked pixels into the model using the newly obtained pose. Note that, since our method performs a second registration step, the registration takes up to twice as long compared to the approach ignoring dynamics.

D. Carving of Model and Free Space Management

One weakness of TSDF approaches is that they cannot keep track of perceived free space. However, being aware of the previously measured free space is a way to reject dynamic objects, as opposed to detecting them from their motion. We follow a rather simple rule: a voxel that was found to be reliably empty can never contain a static object. Indeed, if a new range image points to a voxel being occupied, it can only be so because a dynamic object has entered it. Therefore, we can safely reject it. Doing so, we sidestep the notoriously difficult problem of tracking dynamic moving points, as done in [12], [13]. We mark as free every non-occluded voxel in the camera frustum outside the truncation region (up to a clipping plane). Note that this is particularly helpful in case of measurements of points that are too far from the sensor to be integrated into the TSDF. In our current implementation, we assign to free voxels an SDF value equal to the truncation distance τ . Doing this for every voxel in the frustum is suboptimal in terms of memory consumption. Improvement by maintaining an octree-based representation of the free space is left as future work.

Another consideration regarding free space is that if a previously static object moves, its voxels must be removed from the map. This corresponds to the fact that if a voxel previously mapped as static is detected as empty subsequently, this voxel used to contain a dynamic object and should, therefore, be marked as free. This behavior is automatically achieved by the TSDF representation, i.e., we update the SDF stored in the voxel by performing a weighted average between the current value and the truncation distance τ . Therefore, if a voxel is detected as free long enough, it will be reliably marked as free.



Fig. 5: (a) Raw depth from the RGB-D sensor. (b) Our virtual refined depth from 10 adjacent frames. (~0.3 s delay)

E. Limitations and Handling Invalid Measurements

Marking free voxels in the camera frustum is effective as long as we know from the sensor that those regions of space are empty. However, common commercial RGB-D cameras return depth images that contain invalid measurements, i.e., pixels with a value of zero, as exemplarily shown by black pixels in Fig. 5a. These measurements are invalid either because they are out-of-range or because they are not measurable, e.g., because of scattering media, reflecting surfaces, etc. For our approach to work in every possible case, it is necessary to distinguish between out-of-range and non-measurable values. Two methods to handle such cases are possible.

The first one is to make no distinction and not consider zero values, as it is commonly done in other RGB-D mapping approaches. In this case, our approach will work correctly when the whole scene is in the range of the depth sensor. A limitation of our approach is that in case there are out-of-range values, dynamic objects could be incorrectly added to the model, thus affecting its quality.

The second method is to “correct” non-measurable values if possible. To this end, we create a temporary model from n consecutive frames and generate virtual depths from the registered poses. Then, we fill in the original depth images by replacing every zero value with the corresponding value of the virtual depth. In this way, non-measurable values are usually reduced thanks to the multiple observation. The remaining values are assumed to come from out-of-range measurements of the camera and are replaced with a high fixed depth value. Fig. 5b shows an example of a virtual depth image obtained with this technique. However, this solution assumes that nothing appears closer than the minimum range of the depth sensor. This assumption can be removed in case we add an additional sensor, e.g., a simple sonar on top of the RGB-D sensor, that detects whether there are objects too close to the camera. Moreover, a disadvantage of this method is that the actual model is then generated with a delay of n frames, which might be suboptimal for some robotics applications.

In the following experiments, we employ the second option for modeling the sequences of the TUM RGB-D dataset, since the depth images contain out-of-range values. For the sequences of our dataset, we employ the first option, since the recorded depth is always within the valid range of the depth sensor.

TABLE I: Parameters of our approach used in all experiments.

| Parameter | Value |
|------------------------------------|--------|
| Voxel size | 0.01 m |
| Truncation distance τ | 0.1 m |
| Weight w_c | 0.025 |
| Floodfill threshold θ | 0.007 |
| Residual threshold weight γ | 0.5 |

IV. EXPERIMENTAL EVALUATION

The main contribution of this work is a TSDF-based mapping approach that is able to operate in environments with the presence of highly dynamic elements by relying solely on geometric information, i.e., our approach is completely class agnostic and does not require tracking of objects. Our experiments show the capabilities of our method and support our key claims, which are: (i) our approach is robust to dynamic elements in the environment and provides a camera tracking performance on par or better than state-of-the-art dense SLAM approaches, and (ii) provides a dense 3D model that contains only the static parts of the environment, which is more accurate than other state-of-the-art approaches when compared to the ground truth model.

We provide comparisons with StaticFusion (SF) [14], DynaSLAM (DS) [1] and MaskFusion (MF) [13]. As our approach does not rely on deep neural networks, we make a distinction in our comparison between the pure geometric approach of DynaSLAM (G) and the combined deep neural network+geometric approach (N+G). We tested all approaches on the dynamic scenes of the TUM RGB-D dataset [17], as well as on our dataset, designed to contain highly dynamic scenes. We obtained the reported results by using the open source implementations available for the different approaches, with the exception of MaskFusion, where we only report results from the paper [13].

In all experiments, we used the default parameters provided by the open source implementations and the same holds for our approach, see Tab. I for details. Our default parameters have been determined empirically by trial and error, but similar values gave comparable results.

In the presented tables, we separate the approaches that rely solely on geometric information, from approaches that rely also on neural networks. We highlight in bold the best result among the first category of approaches, as we focus mainly at class agnostic approaches.

A. Performance on TUM RGB-D Dataset

The first experiment shows the performance of our approach with the TUM RGB-D dataset [17].

Note that, since the depth information from these sequences contains out-of-range values, we used the approach described in Sec. III-E to obtain a refined depth, with the temporary model created from $n = 10$ consecutive frames, which corresponds to a model delayed approximately 0.3 s.

Tab. II shows the results of all considered approaches on six sequences of the TUM dataset. From this table, it is clear that DynaSLAM outperforms the other methods. However, DynaSLAM is a feature-based approach and, in contrast to



Fig. 6: Final mesh obtained using our approach on the *walking_static* sequence, in which two people are continuously walking through the scene.

TABLE II: Absolute Trajectory Error (RMS) [m] on dynamic scenes of TUM dataset.

| | Ours | SF | DS (G) | | DS (N+G) | MF |
|--------------------|--------------|-------|--------------|--|----------|-------|
| Dense approach | ✓ | ✓ | ✗ | | ✗ | ✓ |
| sitting_static | 0.009 | 0.014 | 0.009 | | 0.007 | 0.021 |
| sitting_xyz | 0.040 | 0.039 | 0.009 | | 0.015 | 0.031 |
| sitting_halfsphere | 0.110 | 0.041 | 0.017 | | 0.028 | 0.052 |
| walking_static | 0.017 | 0.015 | 0.014 | | 0.007 | 0.035 |
| walking_xyz | 0.099 | 0.093 | 0.085 | | 0.017 | 0.104 |
| walking_halfsphere | 0.104 | 0.681 | 0.084 | | 0.026 | 0.106 |
| Max | 0.110 | 0.681 | 0.085 | | 0.028 | 0.106 |

the other approaches, does not provide a dense model. The three dense mapping approaches show in most of the cases similar results, except for the sequence *walking_halfsphere*, where StaticFusion lost track due to the excess of dynamic elements at the beginning of the sequence, and the sequence *sitting_halfsphere*, where our approach shows worse performance. In terms of 3D reconstruction, our approach is always able to create a consistent mesh of the environment, see Fig. 6 for an example.

The only case where the model built by our approach shows artifacts is on the *walking_xyz* sequence, where a person remains in the model, see Fig. 7. This happens because the person is tracked by the cameraman at the beginning of the sequence and the location where the person stops is never revisited again. Therefore, the algorithm cannot know that the voxels in that location are actually free. This is confirmed by Fig. 8, which shows the relative position error versus the elapsed time. It is evident from the figure that in the first four seconds, i.e., when the camera tracks the person, the error is particularly high. In sum, we are on par with state-of-the-art dense mapping approaches in terms of tracking, but we use a completely different technique based on TSDF instead of surfels.

B. Performance on Bonn RGB-D Dynamic Dataset

The second set of experiments consists of the comparison between the algorithms on our dataset. Our dataset includes 24 highly dynamic scenes, where people perform different tasks, such as manipulating boxes or playing with balloons, see Fig. 9 for some example RGB frames. These tasks often obstruct the camera, creating particularly challenging situation for mapping approaches. We recorded the dataset using an ASUS Xtion Pro LIVE sensor, combined with an Optitrack Prime 13 motion capture system for the ground



Fig. 7: Final mesh obtained using our approach on the *walking_xyz* sequence. In this sequence, the camera follows the person on the right at the beginning and then never revisits that location. Therefore, the person is added in the model.

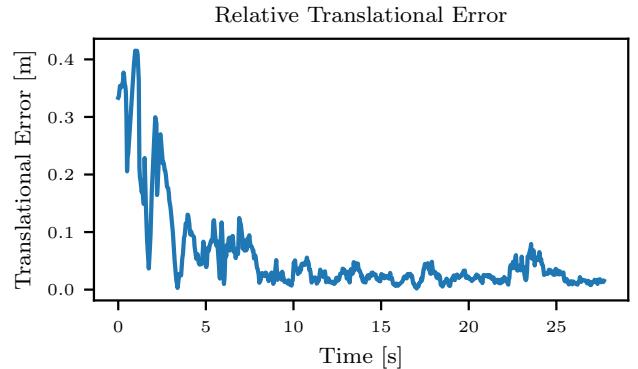


Fig. 8: Relative translational error over time for the *walking_xyz* sequence. In this plot it is visible how the relative error is particularly high at the beginning of the sequence, when the camera is tracking the person. After the first four seconds, the error drops substantially.

truth trajectories. Additionally a Leica BLK360 terrestrial laser scanner was used to obtain a ground truth 3D pointcloud of the static environment.

Tab. III shows the performance of different approaches on our scenes. The variety of sequences show interesting phenomena. For example, on the scenes where the dynamic component is a uniformly colored balloon, DynaSLAM outperforms the dense approaches, because it cannot detect features on the balloon, which therefore does not affect the SLAM performance. Our approach performs best on scenes crowded with people, which are among the most challenging if no semantic segmentation algorithm is available. On sequences that involve the manipulation of boxes, the algorithms have mixed results, with our approach being better in about half of the cases and DynaSLAM being better on the other half. Note that DynaSLAM with the combined neural network and geometric approach performs the best in most cases. This is due to the heavy bias of having people in every sequence of our dataset, therefore the segmentation of people always helps the algorithm achieving better results. However, the worst performance of our approach is on par with the worst performance of DynaSLAM (N+G) and substantially better than the other geometric approaches, showing that our approach is more robust to failure. This is a desirable quality when deploying robotic systems. Finally, as discussed in the previous section, our approach is able to build a consistent model of the environment in most of the cases, see Fig. 1 for an example of model from the scene *crowd3*.



Fig. 9: Example RGB frames from our highly dynamic dataset.

TABLE III: Absolute Trajectory Error (RMS) [m] on our dataset. In this table, we shorten *obstructing_box* with *o_box* and *nonobstructing_box* with *no_box*.

| | Ours | SF | DS (G) | DS (N+G) |
|-------------------|--------------|--------------|--------------|----------|
| Dense approach | ✓ | ✓ | ✗ | ✗ |
| balloon | 0.175 | 0.233 | 0.050 | 0.030 |
| balloon2 | 0.254 | 0.293 | 0.142 | 0.029 |
| balloon_tracking | 0.302 | 0.221 | 0.156 | 0.049 |
| balloon_tracking2 | 0.322 | 0.366 | 0.192 | 0.035 |
| crowd | 0.204 | 3.586 | 1.065 | 0.016 |
| crowd2 | 0.155 | 0.215 | 1.217 | 0.031 |
| crowd3 | 0.137 | 0.168 | 0.835 | 0.038 |
| kidnapping_box | 0.148 | 0.336 | 0.026 | 0.029 |
| kidnapping_box2 | 0.161 | 0.263 | 0.033 | 0.035 |
| moving_no_box | 0.071 | 0.141 | 0.317 | 0.232 |
| moving_no_box2 | 0.179 | 0.364 | 0.052 | 0.039 |
| moving_o_box | 0.343 | 0.331 | 0.544 | 0.044 |
| moving_o_box2 | 0.528 | 0.309 | 0.589 | 0.263 |
| person_tracking | 0.289 | 0.484 | 0.714 | 0.061 |
| person_tracking2 | 0.463 | 0.626 | 0.817 | 0.078 |
| placing_no_box | 0.106 | 0.125 | 0.645 | 0.575 |
| placing_no_box2 | 0.141 | 0.177 | 0.027 | 0.021 |
| placing_no_box3 | 0.174 | 0.256 | 0.327 | 0.058 |
| placing_o_box | 0.571 | 0.330 | 0.267 | 0.255 |
| removing_no_box | 0.041 | 0.136 | 0.016 | 0.016 |
| removing_no_box2 | 0.111 | 0.129 | 0.022 | 0.021 |
| removing_o_box | 0.222 | 0.334 | 0.362 | 0.291 |
| synchronous | 0.441 | 0.446 | 0.977 | 0.015 |
| synchronous2 | 0.022 | 0.027 | 0.887 | 0.009 |
| Max | 0.571 | 3.586 | 1.217 | 0.575 |

C. Model Accuracy

The last set of experiments shows that our approach provides an accurate, dense 3D model that contains only the static parts of the environment.

To perform such experiments, we first built a high resolution point cloud of the static part of our test environment (Fig. 12a) using a professional terrestrial laser scanner, the Leica BLK360 (Fig. 12b). We then aligned the point cloud to our motion capture system’s reference frame using tilt and turn targets (Fig. 12c) that we located with both the laser scanner and the motion capture system. Fig. 10a shows a section of our ground truth point cloud.

To align the model created by the algorithms to our ground truth, we transformed it from the reference frame of the RGB-D sensor, to the reference frame of the motion capture system. We aligned the two frames using the calibration setup shown in Fig. 12d, where the markers positioned in the environment were known in both reference frames.

We compare the models built by our algorithm and by StaticFusion [14] for the sequences *crowd3* and *removing_nonobstructing_box* w.r.t. the ground truth. For each point of the evaluated model, we measure its distance from the

ground truth.

For a qualitative impression, Fig. 10 shows the two models of the scene *crowd3* where the points have been colored according to their distance to the closest point in the ground truth model. In Fig. 10c, one can see that some dynamic elements are still present in the final model, represented by the red points highlighted by the arrow. In contrast, the model from our approach does not show such artifacts caused by dynamic objects.

For a quantitative evaluation, Fig. 11 shows the cumulative percentage of points at a certain distance from the ground truth for the models of the two considered sequences. The plots show in both cases that the reconstructed model by our approach is more accurate.

In summary, our evaluation shows that our method is able to robustly track an RGB-D sensor in highly dynamic environments. At the same time, it provides a consistent and accurate model of the static part of the environment.

V. DATASET AND SOURCE CODE

Our Bonn RGB-D dynamic dataset is available at the URL: <http://www.ipb.uni-bonn.de/data/rgbd-dynamic-dataset>. The source code of our approach is available at the URL: <https://github.com/PRBonn/refusion>.

VI. CONCLUSION

We presented ReFusion: a TSDF-based mapping approach able to track the pose of the camera in dynamic environments and build a consistent 3D model of the static world. Our approach tracks the sensor by exploiting directly the TSDF information and the color information encoded in voxel blocks that are only allocated when needed. Our method filters dynamics using an algorithm based on the residuals from the registration and the representation of free space. We evaluated our approach on the popular TUM RGB-D dataset, as well as on our Bonn RGB-D dynamic dataset, and provided comparisons to other state-of-the-art techniques. Our experiments show that our approach leads to an improved pose estimation in the presence of dynamic elements in the environment, compared to other state-of-the-art dense SLAM approaches. Finally, we publicly release our own dataset, as well as our open-source implementation of the approach.

ACKNOWLEDGMENTS

We thank Raluca Scona for the assistance in running StaticFusion and Berta Bescos for the assistance in running DynaSLAM. Furthermore, we thank Jannik Janßen, Xieyanli Chen, Nived Chebrolu, Igor Bogoslavskyi, Julio Pastrana, Peeyush Kumar, Lorenzo Nardi, and Olga Vysotska for supporting the Bonn RGB-D dynamic dataset acquisition.

REFERENCES

- [1] B. Bescos, J.M. Fcil, J. Civera, and J. Neira. DynaSLAM: Tracking, Mapping, and Inpainting in Dynamic Scenes. *IEEE Robotics and Automation Letters (RA-L)*, 3(4):4076–4083, 2018.
- [2] E. Bylow, J. Sturm, C. Kerl, F. Kahl, and D. Cremers. Real-Time Camera Tracking and 3D Reconstruction Using Signed Distance Functions. In *Proc. of Robotics: Science and Systems (RSS)*, volume 2, 2013.

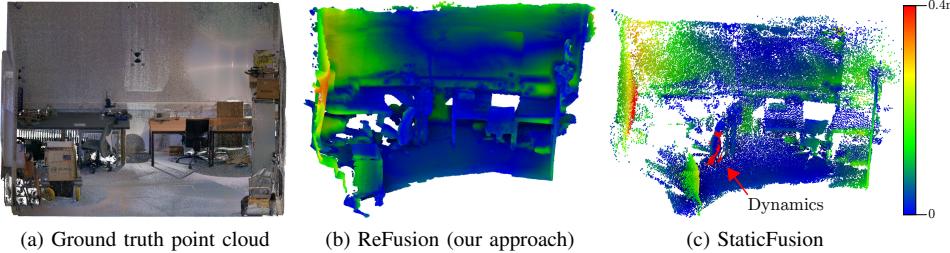


Fig. 10: Models from our approach and StaticFusion of the scene *crowd3* compared against the ground truth. The points of the models are colored according to their distance from the ground truth. The arrow highlights the dynamic parts of the scene still present in the model from StaticFusion

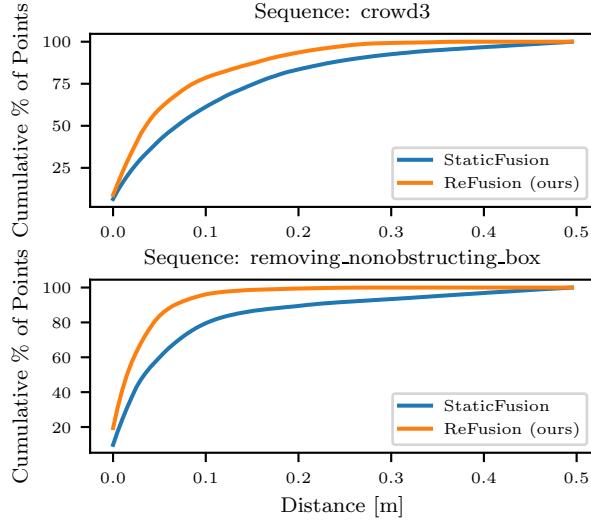


Fig. 11: Plot of the cumulative percentage of points (y axis) at a specific distance from the ground truth 3D model (x axis). The higher the percentage of points towards a zero distance the better.

- [3] D. Canelhas, T. Stoyanov, and A. Lilienthal. SDF tracker: A parallel algorithm for on-line pose estimation and scene reconstruction from depth images. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 3671–3676, 2013.
- [4] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *Proc. of the Int. Conf. on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 303–312. ACM, 1996.
- [5] B. Della Corte, I. Bogoslavskyi, C. Stachniss, and G. Grisetti. A General Framework for Flexible Multi-Cue Photometric Point Cloud Registration. In *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2018.
- [6] O. Kähler, V. Prisacariu, J. Valentin, and D. Murray. Hierarchical Voxel Block Hashing for Efficient Integration of Depth Images. In *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2016.
- [7] M. Keller, D. Lefloch, M. Lambers, and S. Izadi. Real-time 3D Reconstruction in Dynamic Scenes using Point-based Fusion. In *Proc. of the Intl. Conf. on 3D Vision (3DV)*, pages 1–8, 2013.
- [8] R. Mur-Artal and J.D. Tardós. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Trans. on Robotics (TRO)*, 2017.
- [9] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-Time Dense Surface Mapping and Tracking. In *Proc. of the Intl. Symposium on Mixed and Augmented Reality (ISMAR)*, pages 127–136, 2011.
- [10] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3D Reconstruction at Scale using Voxel Hashing. *Proc. of the SIGGRAPH Asia*, 32(6), 2013.
- [11] P.O. Pinheiro, T. Lin, R. Collobert, and P. Dollár. Learning to Refine Object Segments. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, pages 75–91, 2016.
- [12] M. Runz and L. Agapito. Co-Fusion: Real-Time Segmentation, Tracking and Fusion of Multiple Objects. In *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2017.
- [13] M. Rünz, M. Buffier, and L. Agapito. MaskFusion: Real-Time Recognition, Tracking and Reconstruction of Multiple Moving Objects. In *Proc. of the Intl. Symposium on Mixed and Augmented Reality (ISMAR)*, pages 10–20, 2018.
- [14] R. Scona, M. Jaimez, Y.R. Petillot, M. Fallon, and D. Cremers. Staticfusion: Background reconstruction for dense rgbd slam in dynamic environments. In *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2018.
- [15] C. Stachniss, J. Leonard, and S. Thrun. *Springer Handbook of Robotics*. 2nd edition, chapter Chapt. 46: Simultaneous Localization and Mapping. Springer Verlag, 2016.
- [16] F. Steinbrücker, J. Sturm, and D. Cremers. Volumetric 3D Mapping in Real-Time on a CPU. In *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2014.
- [17] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A Benchmark for the Evaluation of RGB-D SLAM Systems. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2012.
- [18] A.S. Vempati, I. Gilitschenski, J. Nieto, P. Beardsley, and R. Siegwart. Onboard Real-time Dense Reconstruction of Large-scale Environments for UAV. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2017.
- [19] T. Whelan, S. Leutenegger, R. S. Moreno, B. Glocker, and A. Davison. ElasticFusion: Dense SLAM Without A Pose Graph. In *Proc. of Robotics: Science and Systems (RSS)*, 2015.
- [20] M. Zollhöfer, P. Stotko, A. Görlitz, C. Theobalt, M. Nießner, R. Klein, and A. Kolb. State of the Art on 3D Reconstruction with RGB-

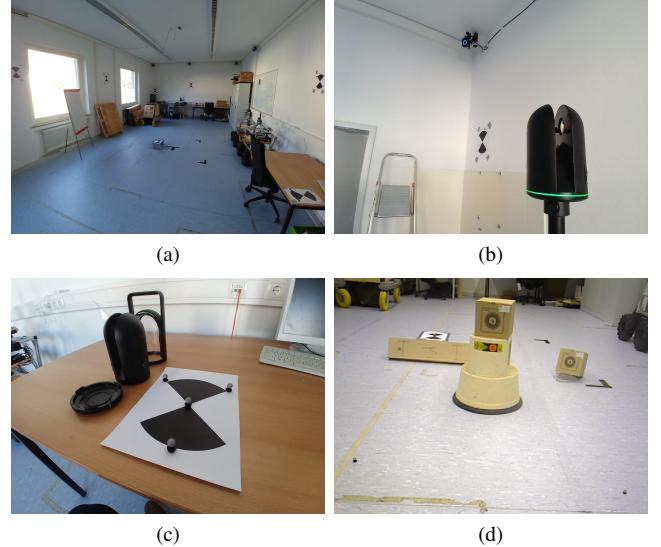


Fig. 12: (a) Our test environment. (b) Terrestrial laser scanner. (c) Tilt and turn target. (d) Calibration setup used to align the sensor’s reference frame with the motion capture system’s one.

D Cameras. In *Eurographics - State-of-the-Art Reports (STARs)*, volume 37, 2018.