Egocentric RGB-D-Thermal: A New Framework for Understanding Human-Object Interactions

Anonymous ICCV submission

Paper ID 1455

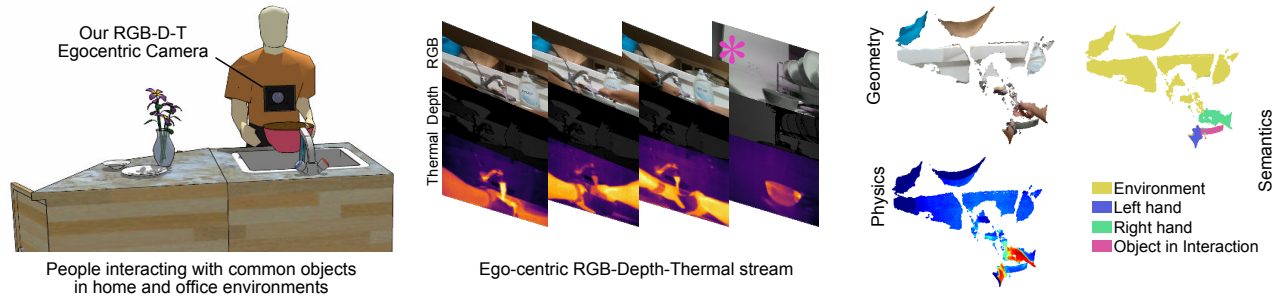


Figure 1: Understanding Human-Object Interactions. We collect a large dataset of egocentric multi-modal videos of humans performing daily activities such as washing dishes. We use RGB, depth, and thermal information as modalities to extract semantics, geometry, and physics information about the scene. We further propose a method for egocentric SLAM with semantics. Consider the right-most frame (labelled with *). The RGB and depth signals do not carry any physics-related information, but the thermal residue on both sides of the bowl clearly suggest that a person just carried that bowl, holding it on both sides with both hands after washing it.

Abstract

Understanding human-object interactions is crucial for both robotics and computer vision. Gaining this understanding requires a combined knowledge of the semantics, geometry, and physics (i.e. force, material properties, etc.) of the interaction; however, these three aspects have never been studied simultaneously. We propose a new dataset and a framework that allows us to jointly consider semantics, geometry, and physics, and thus enables new insights about human-object interactions. Our dataset takes the form of multi-modal RGB, depth, and thermal egocentric videos of everyday activities. To inject structure into this data, we introduce a framework for egocentric SLAM that enables the construction of geometric information for our dataset in the form of 3D maps, as well as a set of camera poses and semantic segmentations. Thus, our framework can be used to learn semantics, geometry, and physics jointly.

1. Introduction

Consider a typical robotics scenario, in which a robot must understand humans, objects, and most importantly the interactions between them. These robots will need to detect, track, and predict human motions [32, 40, 43], and relate them to the objects in the environment. For example, a kitchen robot helping a chef should understand which object the chef is reaching for on a very crowded and highly-occluded kitchen table in order to deduce and bring back the missing next ingredient from the refrigerator. This type of high-level reasoning requires rich semantics for humans, objects, and their **interactions**. This crucial problem of understanding human-object interactions has been widely studied; however, these studies have not translated into real-world robotic interaction and manipulation abilities. We believe that this is largely due to the fact that existing sensing modalities cannot capture the complexity of human-object interactions. We propose to resolve this problem by introducing a new data modality, thermal, together with a framework for making sense of this new multi-modal data.

Successfully understanding human-object interactions requires the joint consideration of three aspects: the semantics of the scene, the geometry of the scene, and the physics of the interaction. For instance, consider an interaction as simple as washing and moving a bowl, shown in Figure 1. Looking at the right-most frame (marked by *), neither the RGB nor the depth image provides information about the interaction. But looking at the thermal image, we can see a thermal residue on both sides of the bowl and conclude that a person carried the bowl by holding it from both sides. In this work, we will fill the gap left by existing sensing methods by introducing a new thermal modality along with an efficient way to collect and structure multi-modal data.

We choose three critical modalities - RGB, depth, and thermal - to include in our raw data in order to understand human-object interactions. We design an affordable hardware to capture all of these modalities by mounting a structured-light camera (RGB-D) and a mobile thermal camera (RGB-Thermal) to a chest harness (Figure 2). We then develop a system to calibrate and time-synchronize these cameras. The resulting data is a 2D stream of RGB, depth, and thermal information.

Although a stream of RGB, depth, and thermal images includes the necessary semantic, geometric, and physical information, it is not structured in a way that is useful for learning about human-object interactions. We propose a framework that can jointly infer the geometry (camera poses), the semantics (human-vs-static environment-vs-object in interaction), and the physics (thermal residue and its temporal decay pattern) of the scene. The problem of jointly inferring semantics and scene geometry can be considered a form of a semantic SLAM problem. After studying the distinct properties of our egocentric setup, we introduce a method for egocentric SLAM with semantics.

In summary, our contributions are: i) Designing an affordable, multi-modal data acquisition system to understand human-object interactions over semantics, geometry, and physics. ii) Sharing a large-scale dataset with annotations of semantic entities (hands and objects) as well as all of our source code and hardware designs for reproducible and effective research on human-object interactions. iii) An egocentric SLAM algorithm that can combine all data modalities, generating environment, human motion, and object tuples that represent human-object interactions to near completion. We envision our proposed real-time SLAM architecture to be useful beyond dataset creation, becoming part of the standard pipeline for human-robot interaction.

2. Related Work

Human-Object Interactions: Numerous attempts have been made over the past several decades to better understand human-object interactions. J.J. Gibson coined the term "affordance" as early as 1977 to describe the action possibilities latent in the environment [7]. Donald Norman later appropriated the term to refer to only the action possibilites that are perceivable by an individual [28].

More recently, [16] and [15] learn human activities by using object affordances. [33] teaches a robot about the world by having it physically interact with objects. [23] predicts long-term sequential movements caused by applying external forces to an object, which requires reasoning about the physics of the situation.

Several works also examine human-object interactions as they relate to hands or grasps. For instance, [38] and [1] both explore grasp classification in an attempt to understand hand-object manipulation. [10] studies the hands to discover a taxonomy of grasp types using egocentric videos.

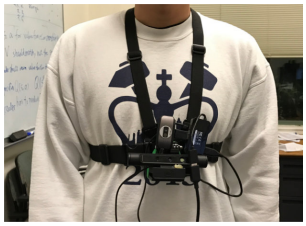
Hand Tracking: Hand tracking has been a particular focus for human-object interactions. [19] and [18] perform pixel-wise hand detection for egocentric videos by using a dataset of labeled hands, and by posing the hand detection problem as a model recommendation task, respectively. [42] and [36] perform depth-based hand pose estimation from a third-person and an egocentric perspective. [41] simultaneously tracks both hands manipulating an object as well as the object pose using RGB-D videos. The closest to our work is [36]; however, it considers only images, lacks thermal information, and experiments only at small scale.

Egocentric Scenes: A few works have also focused on egocentric scenes. For example, [37], [22],[13], and [20] look at first-person pose and activity recognition. [17] creates object-driven summaries for egocentric videos.

SLAM: SLAM is the problem of constructing a map of an unknown environment while tracking the location of a moving camera within it. Although there are visual odometry approaches [12, 21], explicit models of the map typically increase the accuracy of ego-motion estimation as well. Thus, SLAM has become an increasingly popular area of research, especially for robotics or virtual reality applications even when only the ego-motion is needed. Several early papers propose methods for monocular SLAM [3, 9]. More recently, ORB-SLAM proposes a sparse, feature-based monocular SLAM system [24]. LSD-SLAM is a dense, direct monocular SLAM algorithm for large-scale environments [6], and DSO is a sparse, direct visual odometry formulation [5].

Several stereo SLAM methods also exist for RGB-D settings, for example the dense visual method DVO-SLAM [14]. Kintinuous is another dense visual SLAM system that can produce large-scale reconstructions [44]. ElasticFusion is a dense visual SLAM system for room scale environments [45]. ORB-SLAM2 extends ORB-SLAM for monocular, stereo, and RGB-D cameras [25]. KinectFusion can map indoor scenes in variable lighting conditions [27], and BundleFusion estimates globally optimized poses [2].

All of the above algorithms are designed for static scenes from a more global perspective. Nevertheless, although most SLAM systems assume a static environment, a few methods have been developed with dynamic objects in mind. [39] builds a system that allows a user to rotate an object by hand and see a continuously-updated model. [46] presents a structured light technique that can generate a scan of a moving object by projecting a stripe pattern. More recently, DynamicFusion builds a dense SLAM system for reconstructing non-rigidly deforming scenes in real time with RGB-D information [26]. However, these methods reconstruct only single objects rather than entire scenes, and none consider the egocentric perspective.


Camera During Data Collection


Developed Camera Setup

Figure 2. Developed multi-modal camera setup. We combined an RGB-D camera (Intel RealSense SR300) with a mobile thermal camera (Flir One) and attached both to a GoPro chest harness.

3. Dataset

In order to better understand human-object interactions over their semantics, geometry, physics, and appearance, we designed a multi-modal data acquisition system combining an RGB-D camera with a mobile thermal camera. We then used this setup to collect a large dataset of aligned multi-modal videos, and annotated semantically relevant information in these videos. In this section, we explain our process in detail and discuss the characteristics of the collected dataset. In Section 3.1, we describe the hardware that we used; in Section 3.2, we describe the annotations; and in Sections 3.3 and 3.4, we discuss the collected data.

3.1. Hardware

Our data acquisition system includes two mobile cameras: one RGB-D (an Intel RealSense SR300 [34]) and one thermal (a Flir One Android [31]). We mounted both cameras on a GoPro chest harness and connected them through a single USB 3.0 cable to a lightweight laptop kept in the backpack of the data collector. We developed a GNU/Unix driver for the Flir One, since the camera was originally designed for Android mobile phones. We also time synchronized the cameras, using the frame rate of the slower of the two cameras (the Flir One), resulting in a data acquisition rate of 8.33 FPS. We geometrically calibrated the two cameras as explained in Supplementary Materials. The developed camera setup and chest mount are shown in Figure 2. After calibrating, we considered only the spatial area covered by both cameras and saved per-pixel RGB, depth, and temperature values. We relied on RGB values provided by the RGB-D camera and used nearest neighbor interpolation for pixel values that were missing due to the resolution mismatch between different modalities.

3.2. Annotations

The hands and the objects that the hands interact with comprise the key semantic information that we need. We annotated the location of each hand and each object in interaction for each video frame. Since the RGB-D and thermal cameras are jointly calibrated, annotating the the locations in the RGB modality suffices; hence, annotations were done for the RGB channel of the RGB-D camera. Our hand and object annotations take the form of 2D bounding boxes. We also annotated a few segmentations for the hands and objects in interaction.

To obtain the ground truth camera poses, we chose approximately 10 uniformly distributed frames for 10 randomly chosen videos and manually provided a set of corresponding points in each pair of consecutive frames to calculate the relative camera pose. We used least square estimation with RANSAC using the camera calibration matrix provided by Intel.

3.3. Statistics and Visual Examples

Our dataset includes 250 videos of humans performing various activities, as tabulated in Table 1. The activities are divided into four main categories: kitchen, office, household, and recreation. Videos are 3 minutes long on average, and there are over 450,000 frames in total. We observed 20 environments over the dataset, collected by 10 people.

We provide visual examples for some of the activities in Figure 3. One key property of our dataset is that all interactions are very natural, since we did not give the data collectors any specific instructions other than asking them to wear the camera while performing the high-level activity. Because we performed the data collection in various environments, our resulting dataset is also highly diverse in terms of appearance, shape, and interactions.

3.4. Physics

In this section, we qualitatively discuss the physical properties that are captured in our proposed dataset. One interesting phenomenon is that an object that interacts with human hands usually gets warmer since the hand is warm. After the interaction is over, part of the object stays warmer when compared with rest. We refer to this heat imprint from the hand as *thermal residue*, and we believe it acts as an

Table 1. Distribution of collected videos over high level activities.

Activity	Number of Videos
making coffee	12
using microwave	13
using cooking tools	20
food preparation	35
dining	13
cleaning the kitchen	10
studying	33
using office supplies	30
using a computer	10
folding clothes	16
cleaning at home	12
other chores	12
gaming	13
sports	10
art	11

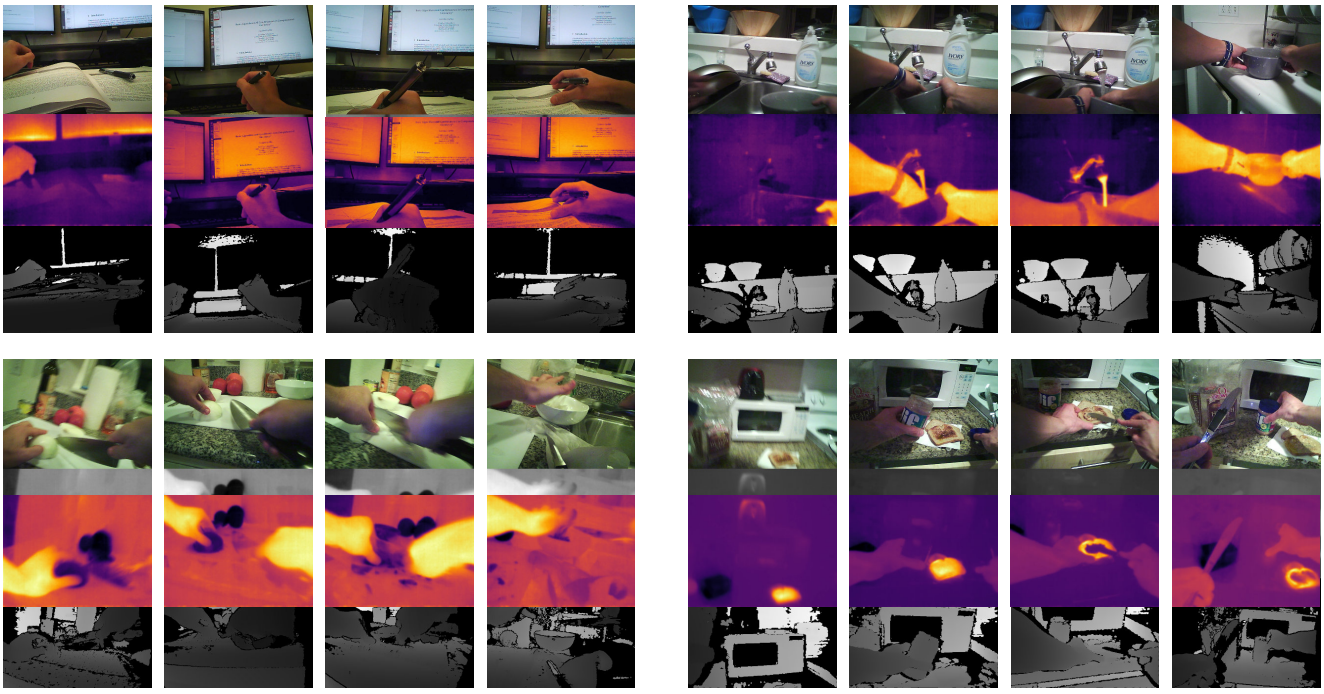


Figure 3. Visual examples from the dataset. We visualize various kitchen and office activities in this figure. Our dataset consists of subjects executing daily activities while wearing our cameras. Because we do not give any specific instructions to the subjects, all motions are natural. Actions are also performed in very different environments for high variability in the dataset.

Table 2. Hand detection accuracy. Our results suggest that thermal and RGB are the most important modalities.

Algorithm	Average Precision (AP)
FORTH [30] (rgb-d)	51.1
Deep Hand [29] (rgb-d)	68.7
YOLO [35]	66.3
YOLO* (rgb-d)	73.6
YOLO* (rgb-t)	86.3
YOLO* (full)	89.2

important indicator of many physical properties including where the hand touches the object, the amount of force applied to the object, and the object material. In Figures 4 and 5, we visualize the behavior of the thermal residue through time for differing amounts of force and for different materials. As the figures suggest, the temporal dissolution pattern of the thermal residue indicates a correlation with force and material properties. For example, in Figure 5, the wood and metal behave very differently because of their different thermal conductivity values. We see very minimal thermal residue on the metallic surface for only a very short amount of time; however, on the wooden surface we see a thermal residue even after 6 seconds. In Figure 4, we see the effect of applying different amounts of force on the thermal residue. We observe the same object (a book) in different activity scenarios, and clearly carrying a book requires more force than flipping through a book. Hence, a stronger force results in a longer and stronger thermal residue. We can conclude that the collected thermal information has considerable potential for capturing and understanding the physics of human-object interactions.

4. Egocentric SLAM with Semantics

In order to fully understand human-object interactions, it is necessary to have a joint knowledge of the humans, the objects in interaction, and their environment. Moreover, this knowledge must include semantics, geometry, and physics to completely describe all aspects of the human-object interaction. As discussed in Section 3, the multimodal dataset that we collected includes all of the required information in an unstructured form. In this section, we explain our proposed method, which converts these raw videos into structured information: we learn the semantics of the scene by segmenting the hands, the objects interacting with the hands, and the static environment points; and we infer the geometry of the scene by learning the camera pose at each timestep and creating a 3D map of the scene.

There are two key problems that need to be solved to achieve this: i) labelling each 3D point with one of the semantic classes: *left hand, right hand, object in interaction, static environment*, and ii) aligning all frames in space in order to construct the geometric information. Since our videos are egocentric, the left and right hands describe the human component, while the object in interaction and the static en-

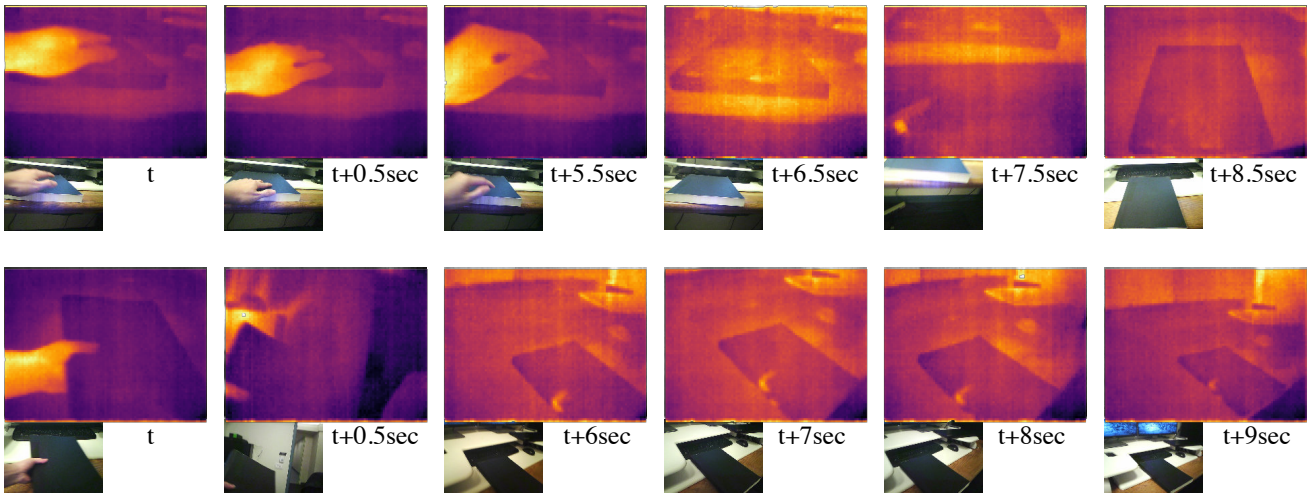


Figure 4. Correlation between thermal residue and force. We observe the same object in two different activities from our dataset. Carrying a book (bottom) requires more force than flipping through it (top). Hence, a higher force results in longer and stronger thermal residue.

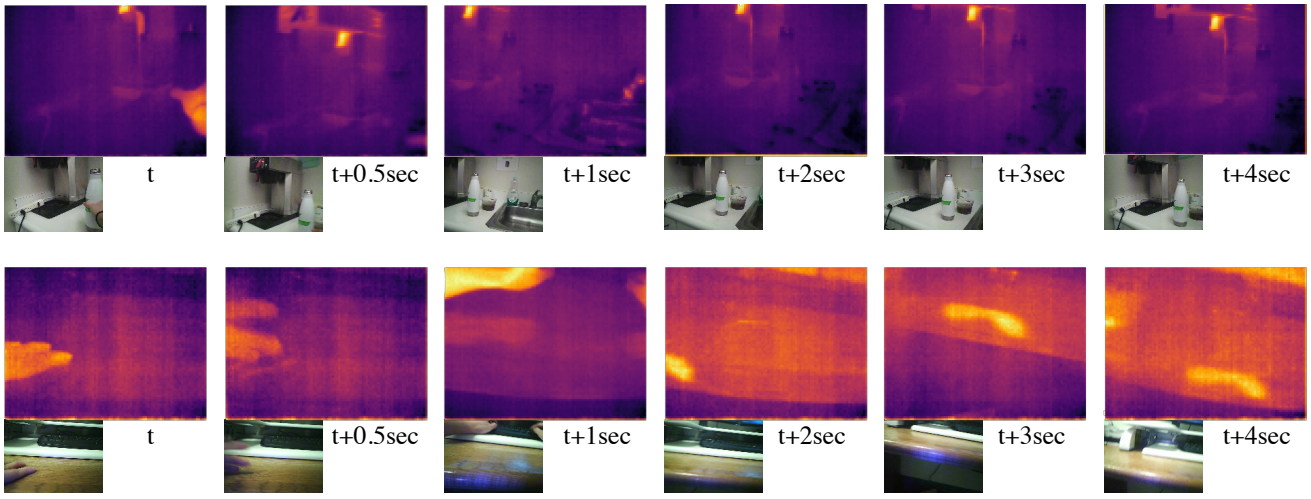


Figure 5. Correlation between thermal residue and material. The responses of the metal water bottle (top) and wooden table (bottom) to temperature are vastly different because of their thermal conductivity values.

vironment describe the remaining components.

This problem can be seen as form of structure-from-motion (SfM) or simultaneous localization and mapping (SLAM) problem with semantics depending on whether the geometric alignment is done in an online or offline fashion. Since our main application area is robotics, we choose to develop an online algorithm. Hence, the problem we address is Egocentric SLAM with semantics. In the remaining part of this section, we first define and discuss the characteristics of the problem of Egocentric SLAM with semantics. Then, we explain our algorithm in detail.

4.1. Problem Definition

Our algorithm takes in an aligned multi-modal video as input and creates a 3D point-cloud over time. Formally, the input to our algorithm is $\{\mathbf{z}_i^t, \mathbf{c}_i^t, d_i^t, \tau_i^t\}_{i\in\{1,\dots,W\times H\}}$ where $\mathbf{z}$ is the 2D-pixel location, $\mathbf{c}$ is the RGB color vector, d is the depth value, and τ is the temperature. W and H are the width and height of the image. The desired output is a dense 3D point cloud as $\{\mathbf{x}_i^t, \mathbf{c}_i^t, \tau_i^t, s_i^t\}_{i\in\{1,\dots,N\}}$ where $\mathbf{x}$ is the 3D location in a global coordinate frame and s is a semantic label as $s \in \{\text{left hand, right hand, object in interaction, static}\}$. We call our problem Egocentric SLAM with semantics since it has distinct properties when compared to a typical SLAM setup. We summarize these distinct properties as follows:

i) Structured Dynamicity: Most SLAM algorithms assume a static environment and cannot handle dynamic objects. The dynamic SLAM methods that do exist are computationally too expensive and can only handle small motions. Our setup is a middle ground between fully static and fully dynamic scenes since all the dynamicity in the scene is caused by the person wearing the camera.

ii) Well defined semantics: In our setup, we need to annotate each point with a well-defined class from *left hand, right hand, object in interaction, static environment*. These

classes are well defined since they are caused by the geometry of the human-object interaction.

iii) High Camera Motion: Our camera is chest-mounted on a human with a harness, so it experiences a large amount of motion due to the non-linear movements of the humans and the elasticity of the harness. This setup is very different from the slowly-panning videos typically used in SLAM.

4.2. Approach

Our method iterates over camera localization, semantic segmentation, sparse mapping, and dense mapping for each frame. Given the multi-modal video frame with color, depth, and temperature information, we start by solving the pose of the camera as $\mathbf{R}^t$ and $\mathbf{t}^t$. Using the camera pose, we label each pixel in the input frame as one of *left hand, right hand, object in interaction, static environment*. Then, we use only the static environment points to update the internal sparse map. We also update the dense map. We keep both a sparse and a dense map for representativeness and robustness. We need a robust 3D map (sparse) for accurate estimation of the camera pose, but we also need a dense map to understand human-object interactions in detail.

Our camera localization and sparse mapping is based on extending an existing SLAM method. We carefully extend ORB-SLAM [25] for our setup since it is robust to camera motion and motion blur. We use the camera localization and sparse mapping submodules of ORB-SLAM [25] by masking our input to include only the static points; this step is necessary since ORB-SLAM handles only static scenes. Our key contribution is extending existing SLAM methods with the knowledge of static and dynamic regions resulting from our accurate and effective semantic segmentation algorithm. We skip the details of ORB-SLAM here and provide a short summary in the supplementary materials for the sake of completeness, as our contributions are general and can be applied to any SLAM algorithm.

4.3. Semantic Segmentation - Dynamic vs. Static

One of the key problems in our egocentric SLAM with semantics framework is the segmentation of the input frame into *left hand, right hand, object in interaction, static environment* classes. The importance of this segmentation is two-fold: 1) removing the dynamic points from the input frame is critical for successful SLAM operation, and 2) these labels create the structure needed to understand human-object interactions.

We perform semantic segmentation after the initial camera localization. Although one can argue that the camera localization step requires static-vs-dynamic segmentation since ORB-SLAM implicitly assumes a static scene, using the entire scene for localization is reasonable since only the the updated sparse map, which includes only static points, is used for the estimation. Having the camera pose is also useful for semantic segmentation because accurate pose is instrumental in reasoning about moving objects.

We have priors for the hands due to their consistent color model, high temperature, and distinct shape. Moreover, hand location is a prior for the location of the object in interaction since the object needs to be in contact with the hand. Therefore, first solving for the hand segmentation and then using it for segmenting the object in interaction is a sound approach. Hence, we perform semantic segmentation as a two-step process: first, we segment the left and right hands from the frame; then, we segment the object in interaction.

We use CRF based image segmentation for this purpose, defining an energy minimization problem:

$$\min_{\alpha_i^t} \sum_i U(\alpha_i^t, \mathbf{y}_i^t) + \sum_i \sum_{j \in \mathcal{N}(i)} V(\mathbf{y}_i^t, \mathbf{y}_j^t) \mathbb{1}[\alpha_i^t \neq \alpha_j^t] \quad (1)$$

In this formulation, α_i^t is a binary variable which is 1 if pixel i is part of the hand at time t and 0 otherwise. $\mathcal{N}(i)$ is the set of pixels neighboring i, and $\mathbb{1}(\cdot)$ is an indicator function. $\mathbf{y}$ is the concatenated vector of $\mathbf{z}, \mathbf{c}, d, \tau$.

$U(\alpha_i^t, \mathbf{y}_i^t)$ is the unary energy representing the likelihood that the i^{th} pixel is part of the hand. It is a weighted combination of the likelihood over the temperature (T), color (C), hand-detector outputs (S), and history over time (H).

$$\begin{aligned} U(\alpha_i^t, \mathbf{y}_i^t) &= w_T U^T(\alpha_i^t, \mathbf{y}_i^t) + w_C U^C(\alpha_i^t, \mathbf{y}_i^t) \\ &+ w_S U^S(\alpha_i^t, \mathbf{y}_i^t) + w_H \sum_i U(\alpha_i^{t-1}, \mathbf{y}_i^{t-1}) e^{-\Delta(\mathbf{y}_i^t, \mathbf{y}_i^{t-1})} \end{aligned} \quad (2)$$

where $\Delta(\cdot,\cdot)$ is the geodesic distance over all modalities between two points *(see Supplementary Materials for formal definition)*. $V(\cdot,\cdot)$ is the binary consistency term defined over neighboring pixels as

$$V(\mathbf{y}_i^t, \mathbf{y}_j^t) = \exp\left(-\frac{|\mathbf{y}_i^t - \mathbf{y}_j^t|_2}{\gamma}\right) \quad (3)$$

where $\gamma = \frac{1}{N} \sum_i \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} |\mathbf{y}_i^t - \mathbf{y}_j^t|_2$, and N is the total number of pixels.

We define the each component of the unary energy as

$$\begin{aligned} U^T(\alpha_i^t, \mathbf{y}_i^t) &= \tau_i^t \mathbb{1}[\alpha_i^t = 1] + (1 - \tau_i^t) \mathbb{1}[\alpha_i^t = 0] \\ U^C(\alpha_i^t, \mathbf{y}_i^t) &= p(\mathbf{c}_i^t | \alpha_i^t) \\ U^S(\alpha_i^t, z_i^t) &= \sum_{k \in \mathcal{H}} p_k e^{-\Delta(\mathbf{y}_i^t, \mathbf{y}_k)} \mathbb{1}[\alpha_i^t = 1] \end{aligned} \quad (4)$$

where $p(\mathbf{c}_i^t | \alpha_i^t)$ is an RGB-color model represented as a Gaussian Mixture Model (GMM) with five components and learned separately for the hand and the static scene. U^S is a collection of hand detections, each represented by a centroid $\mathbf{c}_k$ and a detection likelihood p_k. $\mathbf{y}_k$ is the color, position, depth and temperature of centroid of the detected hand.

All components of this energy function can be computed in log-linear time using bi-linear filters, and minimized using the min-cut/max-flow framework as explained in [47]. We use the open source code released by the authors of [47] and refer the readers to the original paper for details.

After the hands are segmented, we segment the rest of the image into static and dynamic object parts. We use the same energy minimization framework with an additional prior on motion. This prior corresponds to the fact that the motion of the object in interaction is different from the camera motion and is defined as:

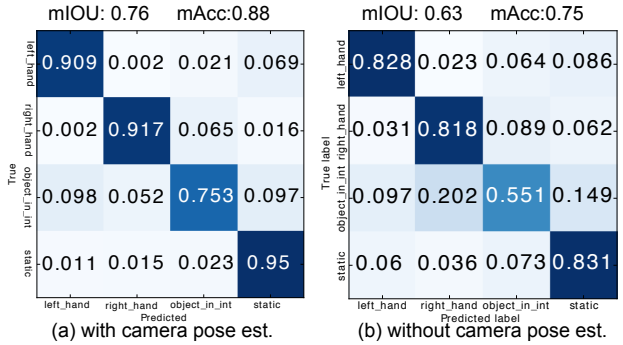
$$U^M(\alpha_i^t, \mathbf{y}_i^t) = \rho(|\mathbf{z}_i^t - \mathbf{z}^{t-1}_{\pi(\mathbf{R}^t\mathbf{x}_i^t + \mathbf{t}^t)}|) \quad (5)$$

where ρ is the Huber function, π is the pinhole projection, $\mathbf{R}, \mathbf{t}$ are the estimated camera pose, and $\mathbf{X}_i$ is the 3D position of i^{th} point in homogeneous coordinates. With some abuse of notation, α_i is a binary variable which is 1 if pixel i is part of the object in interaction and 0 otherwise.

To learn the tradeoff parameters w_T, w_S, w_H, w_M, we use cross-validation and explain the implementation details in Supplementary Materials.

Hand Detection: We use all three modalities to detect the hands and design an algorithm based on the YOLO object detector [35] for real-time performance. Our analysis suggests that 2D bounding box detection using all three modalities results in higher accuracy than state-of-the-art, model-based RGB-D hand pose detection algorithms. In order to train the YOLO detector, we use features pre-trained on ImageNet [4] and train with the annotated bounding boxes in the dataset. Since the pre-training exists only for RGB images, we use knowledge distillation [8] for transferring pre-trained features to thermal and depth modalities. *(See Supplementary Materials for details).*

Dense 3D Map of Static Points: We need a dense 3D map of the scene in order to study human-object interactions. Typically, bundle adjustment type algorithms are used for this purpose; however, our analysis suggests that our camera localization is quite accurate. Thus, we directly transform all frames into the coordinate system of the first frame and overlap all points.



(a) with camera pose est. (b) without camera pose est.

Figure 6. Confusion matrix for semantic segmentation.

5. Experimental Results

We perform several experiments to evaluate our design choices and demonstrate the effectiveness of our algorithm. We start by discussing qualitative results. Our algorithm is designed to generate the semantics, geometry, and physics of human-object interactions. To visualize both the intermediate steps and the final result, in Figure 7 we show the input RGB-D point clouds as well as their semantic segmentations at various time steps for two sequences.

As Figure 7 suggests, our algorithm results in accurate semantic segmentation, and the semantic reasoning results in accurate geometric information after the egocentric SLAM procedure is completed. Consider the input in the right column, for instance. It is difficult to segment even for a human; however, our algorithm accurately segments the hands and the object in interaction (a 3-hole punch).

Semantic Reasoning: We believe that semantic reasoning is key for efficient camera pose estimation. Although there are a few existing methods for SLAM with dynamic objects [26, 11], none of them include open source code and therefore we cannot compare our methods with theirs. Nonetheless, both existing dynamic SLAM papers state that their methods apply only when the objects are large and the motion is very slow - which is certainly not the case for our data. Hence, these methods are not applicable.

In order to quantify the semantic segmentation quality, we annotate test videos with their semantic segmentations and compute the confusion matrix shown in Figure 6. Our confusion matrix suggests that our semantic segmentation is very accurate. In order to evaluate the effect of geometric reasoning, we also compare our semantic segmentation quality against a baseline not using camera pose estimation (we simply ignore the $U^M(\cdot)$ term in segmentation). As the result in Figure 6 suggests, the camera pose estimation is crucial for accurate segmentation. Hence, semantic segmentation and SLAM should be solved jointly.

In order to further evaluate the effect of the semantic segmentation, we compare our egocentric SLAM against a version without semantic segmentation (vanilla ORB-SLAM) in Table 3 for camera localization accuracy. Results suggest that semantic segmentation is an integral part of our pipeline. For example, for the cases of filling a cup and reading, the vanilla ORB-SLAM loses the tracker and fails to produce any output without our semantic segmentation.

Hand Detection: The hand detection is another important part of our algorithm. We quantitatively analyze hand detection performance by comparing various state-of-the-art object detection and RGB-D hand detection algorithms for the hand detection problem, as shown in Table 2. We name our algorithm as YOLO* simlpy representing YOLO[35] combined with cross-modal distillation. We also compare the effect of each modality on the accuracy.

Our results suggest that the thermal information is the

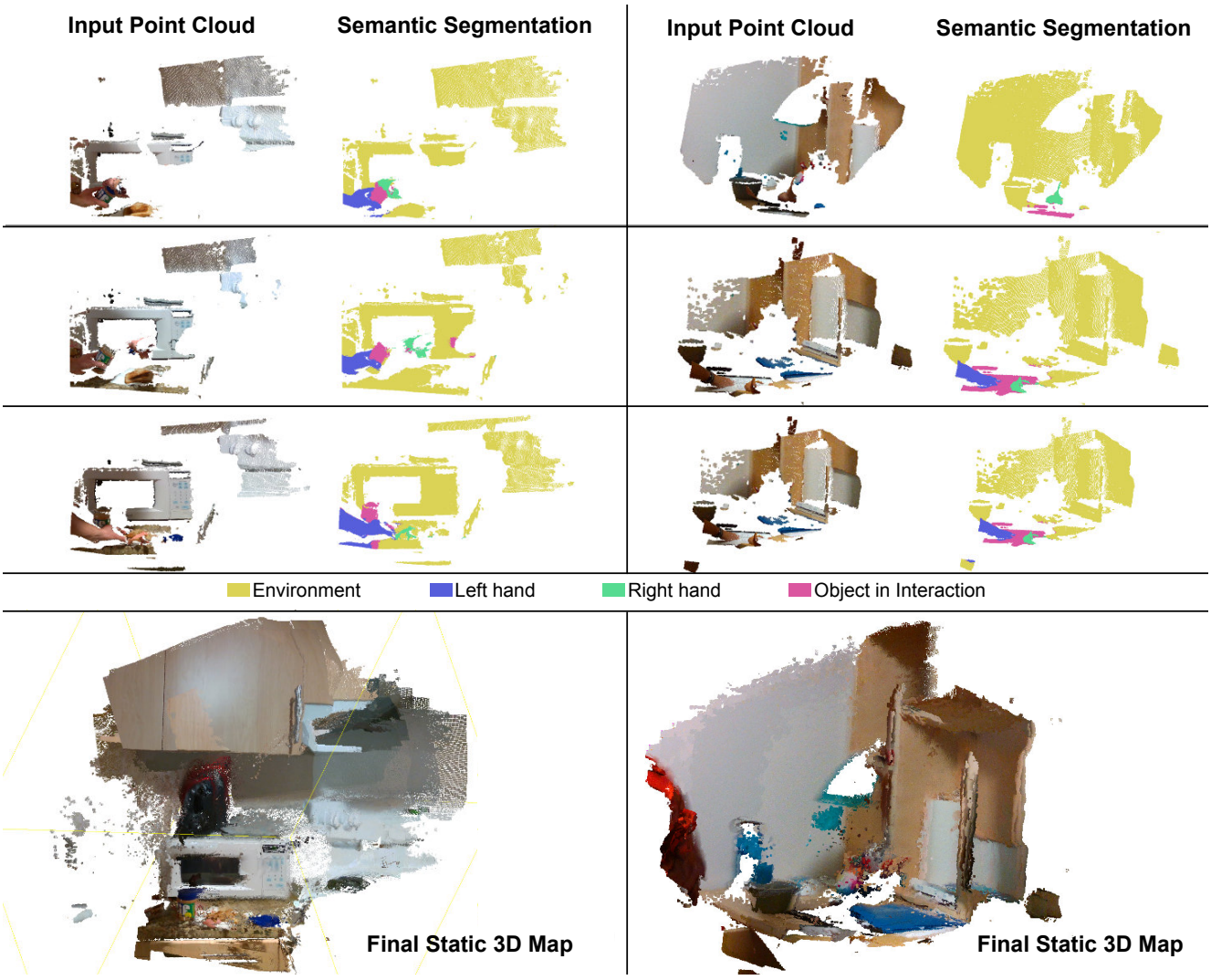


Figure 7. Qualitative results for our egocentric SLAM approach. We show various RGB-D point clouds as input, as well as their semantic segmentations. We also display the final 3D dense static map, reconstructed by combining all input point clouds.

Table 3. Comparison of our method against a version without semantic segmentation. Lost corresponds to videos where SLAM diverged.

		Can opener	Cutting vegetables	Filling a cup	Hole punch	Micro-wave	Peanut butter	Reading	Washing dishes	Average
rMSE Translation Error (cm)	No semantics	4.48	0.81	Lost	4.00	3.48	19.62	Lost	1.78	6.13
	Full method	**4.44**	**0.79**	**1.12**	**3.88**	**2.56**	**12.76**	**1.87**	**1.77**	**3.64**
rMSE Rotation Error ($0.01 \times rad$)	No semantics	4.19	0.96	Lost	2.32	3.81	9.59	Lost	2.20	3.84
	Full method	**4.09**	**0.93**	**0.89**	**2.06**	**3.48**	**7.66**	**1.60**	**2.17**	**2.86**

most useful modality. This result is unsurprising because a human hand is typically warmer than its surroundings. However, a thermal image by itself is not enough since there are usually other warm objects in the environment as well (computer monitors, etc.). Our simple 2D bounding box hand detection outperforms model-based articulated hand pose estimation methods like [30] and [29]; thus, a careful combination of the thermal, RGB, and depth modalities can act as a very powerful framework for hand detection.

6. Conclusion

We present a novel framework that combines the semantics, geometry, and physics of human-object interactions for the first time. Our framework comprises an affordable camera setup, a large-scale dataset of egocentric videos of everyday activities, and an egocentric SLAM algorithm tailored to our requirements. Our results show state-of-the-art performance for hand segmentation, 3D reconstruction, and pose estimation. They also provide rich information about object interactions, ready for use in robotics pipelines.

References

[1] M. Cai, K. M. Kitani, and Y. Sato. Understanding hand-object manipulation with grasp types and object attributes. *Conference on Robotics Science and Systems (RSS)*, 2016. 2

[2] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using online surface re-integration. *arXiv preprint arXiv:1604.01093*, 2016. 2

[3] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. Monoslam: Real-time single camera slam. *IEEE transactions on pattern analysis and machine intelligence*, 29(6), 2007. 2

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 7

[5] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. In *arXiv:1607.02565*, July 2016. 2

[6] J. Engel, T. Schöps, and D. Cremers. Lsd-slam: Large-scale direct monocular slam. In *European Conference on Computer Vision*, pages 834–849. Springer, 2014. 2

[7] J. Gibson James. The theory of affordances. *Perceiving, Acting, and Knowing, Eds. Robert Shaw and John Bransford*, 1977. 2

[8] S. Gupta, J. Hoffman, and J. Malik. Cross modal distillation for supervision transfer. In *In Proc. Computer Vision and Pattern Recognition (CVPR)*, 2016. 7

[9] S. Holmes, G. Klein, and D. W. Murray. A square root unscented kalman filter for visual monoslam. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 3710–3716. IEEE, 2008. 2

[10] D.-A. Huang, M. Ma, W.-C. Ma, and K. M. Kitani. How do we use our hands? discovering a diverse set of common grasps. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[11] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 7

[12] A. Jaegle, S. Phillips, and K. Daniilidis. Fast, robust, continuous monocular egomotion computation. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 773–780. IEEE, 2016. 2

[13] H. Jiang and K. Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. *arXiv preprint arXiv:1603.07763*, 2016. 2

[14] C. Kerl, J. Sturm, and D. Cremers. Dense visual slam for rgb-d cameras. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2100–2106. IEEE, 2013. 2

[15] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013. 2

[16] H. S. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. In *Robotics: Science and Systems (RSS)*, 2013. 2

[17] Y. J. Lee and K. Grauman. Predicting important objects for egocentric video summarization. *International Journal of Computer Vision*, 114(1):38–55, 2015. 2

[18] C. Li and K. M. Kitani. Model recommendation with virtual probes for egocentric hand detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2624–2631, 2013. 2

[19] C. Li and K. M. Kitani. Pixel-level hand detection in egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3570–3577, 2013. 2

[20] M. Ma, H. Fan, and K. M. Kitani. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1894–1903, 2016. 2

[21] Y. Ma, S. Soatto, J. Kosecka, and S. S. Sastry. *An invitation to 3-d vision: from images to geometric models*, volume 26. Springer Science & Business Media, 2012. 2

[22] T. McCandless and K. Grauman. Object-centric spatio-temporal pyramids for egocentric activity recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 2, page 3, 2013. 2

[23] R. Mottaghi, M. Rastegari, A. Gupta, and A. Farhadi. "what happens if..." learning to predict the effect of forces in images. *CoRR*, abs/1603.05600, 2016. 2

[24] M. J. M. M. Mur-Artal, Raúl and J. D. Tardós. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 2

[25] R. Mur-Artal and J. D. Tardós. ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. *arXiv preprint arXiv:1610.06475*, 2016. 2, 6

[26] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015. 2, 7

[27] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. a. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE ISMAR*. IEEE, October 2011. 2

[28] D. A. Norman. *The psychology of everyday things*. Basic books, 1988. 2

[29] M. Oberweger, P. Wohlhart, and V. Lepetit. Hands deep in deep learning for hand pose estimation. *arXiv preprint arXiv:1502.06807*, 2015. 4, 8

[30] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BMVC*, number 2, page 3, 2011. 4, 8

[31] F. One Android. http://www.flir.com/flirone/android/. *Accessed in*, 2017. 3

[32] M. Philipose, K. P. Fishkin, M. Perkowitz, D. J. Patterson, D. Fox, H. Kautz, and D. Hahnel. Inferring activities from interactions with objects. *IEEE Pervasive Computing*, 3(4):50–57, Oct 2004. 1

[33] L. Pinto, D. Gandhi, Y. Han, Y. Park, and A. Gupta. The curious robot: Learning visual representations via physical interactions. *CoRR*, abs/1604.01360, 2016. 2

[34] I. RealSense. www.intel.com/realsense. *Accessed in*, 2017. 3

[35] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. 4, 7

[36] G. Rogez, M. Khademi, J. Supančič III, J. M. M. Montiel, and D. Ramanan. 3d hand pose detection in egocentric rgb-d images. In *Workshop at the European Conference on Computer Vision (ECCV)*, pages 356–371. Springer, 2014. 2

[37] G. Rogez, J. S. Supancic, and D. Ramanan. First-person pose recognition using egocentric workspaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4325–4333, 2015. 2

[38] G. Rogez, J. S. Supancic, and D. Ramanan. Understanding everyday hands in action from rgb-d images. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 3889–3897, Washington, DC, USA, 2015. IEEE Computer Society. 2

[39] S. Rusinkiewicz, O. Hall-Holt, and M. Levoy. Real-time 3d model acquisition. *ACM Trans. Graph.*, 21(3):438–446, July 2002. 2

[40] O. Sener and A. Saxena. rcrf: Recursive belief estimation over crfs in rgb-d activity videos. In *Robotics: Science and Systems*. Citeseer, 2015. 1

[41] S. Sridhar, F. Mueller, M. Zollhoefer, D. Casas, A. Oulasvirta, and C. Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 2

[42] J. S. Supancic, III, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan. Depth-based hand pose estimation: Data, methods, and challenges. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 2

[43] D. K. S. T. Varun Ganapathi, Christian Plagemann. Real-time human pose tracking from range data. In *ECCV*, 2012. 1

[44] T. Whelan, M. Kaess, H. Johannsson, M. Fallon, J. J. Leonard, and J. McDonald. Real-time large scale dense rgb-d slam with volumetric fusion. In *IJRR*, 2014. 2

[45] T. Whelan, S. Leutenegger, R. F. Salas-Moreno, B. Glocker, and A. J. Davison. Elasticfusion: Dense slam without a pose graph. In *RSS*, 2015. 2

[46] L. Zhang, B. Curless, and S. M. Seitz. Rapid shape acquisition using color structured light and multi-pass dynamic programming. In *The 1st IEEE International Symposium on 3D Data Processing, Visualization, and Transmission*, pages 24–36, June 2002. 2

[47] O. ener, K. Ugur, and A. A. Alatan. Efficient mrf energy propagation for video segmentation via bilateral filters. *IEEE Transactions on Multimedia*, 16(5):1292–1302, Aug 2014. 7