

# Self-Supervised Object-in-Gripper Segmentation from Robotic Motions

Wout Boerdijk<sup>1</sup>

Martin Sundermeyer<sup>1</sup>

<sup>1</sup>Institute of Robotics and Mechatronics  
German Aerospace Center (DLR)  
82234 Wessling, Germany  
Firstname.Lastname(at)dlr.de

Maximilian Durner<sup>1</sup>

Rudolph Triebel<sup>1,2</sup>  
<sup>2</sup> Department of Computer Science  
Technical University of Munich (TUM)  
85748 Garching, Germany

**Abstract:** Accurate object segmentation is a crucial task in the context of robotic manipulation. However, creating sufficient annotated training data for neural networks is particularly time consuming and often requires manual labeling. To this end, we propose a simple, yet robust solution for learning to segment unknown objects grasped by a robot. Specifically, we exploit motion and temporal cues in RGB video sequences. Using optical flow estimation we first learn to predict segmentation masks of our given manipulator. Then, these annotations are used in combination with motion cues to automatically distinguish between background, manipulator and unknown, grasped object. In contrast to existing systems our approach is fully self-supervised and independent of precise camera calibration, 3D models or potentially imperfect depth data. We perform a thorough comparison with alternative baselines and approaches from literature. The object masks and views are shown to be suitable training data for segmentation networks that generalize to novel environments and also allow for watertight 3D reconstruction.

**Keywords:** Self-Supervised Learning, Object Segmentation

## 1 Introduction

Despite the major advancements recently made in robotic perception, in practice there are still a lot of challenges to be solved. Two of the most prominent are (1) the lack of annotated training data for variable recognition tasks and (2) domain gaps originating from varying environments and sensors. Techniques like transfer-, semi- and meta- learning can lower the amount of necessary annotations but they do not abstain from it. Another approach is to leverage synthetic data while ensuring generalization to real data through photo-realistic rendering, domain randomization or domain adaptation (e.g. [1, 2]). However, creating the required 3D content is a laborious process and real labeled sensor recordings from environments relevant for execution often yield better results.

In this paper we investigate a different approach: Robots can concurrently observe and interact with objects similar to the way human infants examine them [3]. Concretely, we perform motions of a grasped object in front of a camera and analyze the differences over sequences of images. While humans accomplish discerning objects in motion with specific velocity-sensitive cells [4], we rely on purpose-made Convolutional Neural Networks (CNNs) like LiteFlowNet [5]. These correlate feature maps of subsequent video frames to predict dense optical flow fields. This low-level vision task generalizes well to unseen environments and can be trained efficiently in simulation. To differentiate the resulting joint object-robot mask we refrain from direct object-agnostic segmentation and instead employ a semantic segmentation network to identify the manipulator from static frames in a self-supervised manner. During inference, the trained model removes the robotic arm from the optical flow field to segment the unknown grasped object. Thereby, we automatically obtain accurate masks of any graspable object from any possible viewpoint. These can be used in various downstream tasks like self-supervised semantic or instance segmentation and even full 3D object reconstruction from a grasped object (see Figure 1).

Our method has major advantages. First, it is independent of precise camera calibration, allowing to place an RGB camera loosely pointing towards any robotic manipulator. Second, it does not rely on

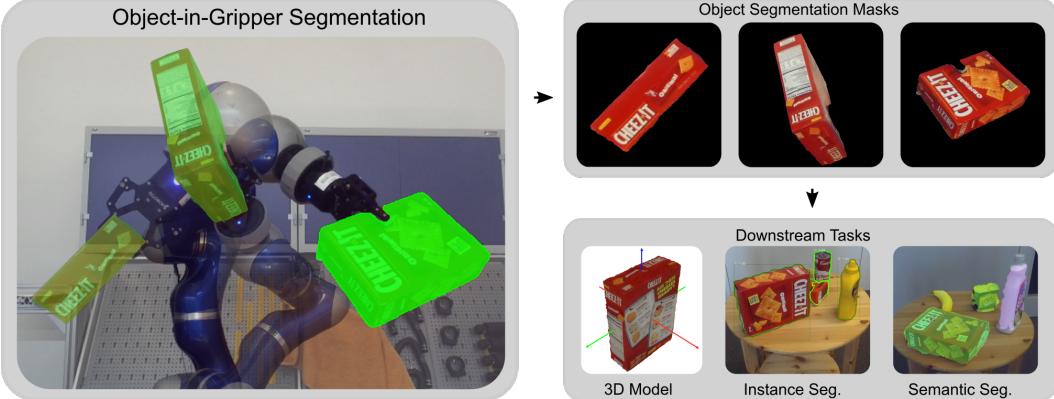


Figure 1: Our self-supervised method is able to segment unknown grasped objects from robotic motions. The resulting annotation masks can be applied to several downstream tasks such as 3D modeling or as training data for instance or semantic segmentation.

3D robot or object models for segmentation. The simple setup enables incremental data collection in varying environments. For the mask generation we only assume that objects are graspable and that the background is mostly static. There are no assumptions on strict object rigidity or texture. Finally, our approach outperforms competing self-supervised methods yielding accurate segmentation performance in novel scenes.

## 2 Related Work

We review existing approaches that induce motion in a scene in order to perform object segmentation. For an extensive analysis on interactive perception we refer the reader to [6].

### 2.1 Object Segmentation from Motion

Motion as an additional cue is a strong enhancement for unknown object segmentation and has widely been investigated. A particularly naïve form thereof is change detection, where subtraction between a pair of images yields the object in motion [7]. More sophisticated methods model the statistical representation of a scene background with, for instance, (multiple) Gaussian distribution(s) or kernel density estimation [8, 9]. Recently, CNNs have also been applied with promising results, and we refer the reader to [10] for an overview.

Spatio-temporal clustering for object segmentation has first been applied by [11]. The main idea is to group pixels based on similar motion behavior obtained by tracking their motion with optical flow over a sequence. Optical flow is a vector field describing the displacement directions and velocities of each pixel between two subsequent images [12]. This underconstrained problem can be solved either by means of energy-based optimization or by employing a learning-based approach [13]. Fragiadaki et al. [14] apply a CNN for object motion detection which is forwarded to a clustering method to obtain object masks. In [15], classical geometric knowledge via optical flow is combined hierarchically with semantic segmentation obtained by a CNN. Recently, two works propose to merge appearance (RGB) and motion (optical flow) information either in a two-stream CNN [16] or in a recurrent neural network [17]. Shao et al. [18] train on synthetic RGB-D frames to estimate object masks as well as a dense 3D motion field, also called scene flow.

These earlier works utilize motion for either binary foreground segmentation or as cue for multi-class supervised segmentation. Instead, we use optical flow to separate two unknown moving instances from each other without requiring any manual annotations.

### 2.2 Robot-based Object Segmentation

In most cases, robots face a static scene which typically does not move by itself. A robot however can enhance visual perception by creating helpful motion in an environment, an idea which has

initially been suggested by [19]. Other works follow this approach by inducing object motion via non-prehensile actions [8, 20]. In [21], generated object hypotheses are verified by pushing, and emerging feature motions are then clustered to object masks. The restriction of textured objects is removed in the subsequent work by applying color annotated stereo data for the hypotheses generation [20]. These probabilities are updated based on optical flow and geometric change detection induced by robotic manipulations. Pathak et al. [22] propose an active agent who learns to segment by trying to pick-and-place objects from one location to another. By observing eventual pixel changes within the grasp region object hypotheses are enforced.

Merely pushing, picking and placing objects falls short in generating diverse object views, a characteristic known to be beneficial for the generalization ability of neural networks and essential for 3d object reconstruction. However, robotic manipulators can also deterministically control the motion of objects in their gripper and thus efficiently generate vast and diverse viewpoints. Deliberately generating trajectories of objects emulates the infant learning process of perceiving objects [4]. Furthermore, the progress in unknown object grasping in recent years [23] encourages us to explore this path to increase robot autonomy.

Most recently, Rocha et al. [24] utilize the kinematic model of a robot for automatic object segmentation in a surgical setting. Given a robot-camera calibration together with a robot model, they project the latter onto the image plane. By iteratively optimizing a GrabCut-based cost function on a set of images followed by projecting the robot arm given the respective kinematics, they obtain segmentation labels. These are used to train a FCN, and a post-processing by a CRF results in tool segmentation. The subsequent work of [25] builds upon this and differs from the aforementioned method by using joint locations in combination with a depth sensor instead of a kinematic robot model projected onto the image frames.

In this work, we propose a novel pipeline to automatically generate training data in a self-supervised setting. Little constraints on the setup and known deep learning components ensure simple practical appliance. In contrast to existing methods we solely use a single RGB sequence for all tasks and do not incorporate depth imagery which is susceptible to errors. Furthermore, we do not require precise camera calibration or registration of object key points in the camera frame. Finally, we explicitly optimize the trajectory in order to maximize the object views.

### 3 Method

In this section, we describe in detail how a combination of optical flow and learned segmentation can be leveraged to automatically segment both, a robotic arm and a grasped object, from single RGB camera streams only. See Figure 2 for an overview.

#### 3.1 Optical Flow Segmentation

We employ a LiteFlowNet [5] trained on synthetic data which is able to efficiently estimate optical flow between two images whilst generalizing well to real environments. By thresholding the velocity magnitude we segment a moving robot arm either with or without a grasped object from static background. Instead of a predefined, hard-coded threshold which would constrain us to a specific velocity or camera resolution we apply Otsu’s binarization [26] on the flow magnitude. In order to counteract under-segmentation of the LiteFlowNet we take the joint mask of thresholded forward ( $t \rightarrow t+1$ ) and backward ( $t \rightarrow t-1$ ) flow to obtain a segmentation mask.

#### 3.2 Visible Robot Arm Segmentation

Still, optical flow alone is unsuitable to distinguish a robot arm from an unknown, grasped object. Therefore, we propose to learn how to segment the robot arm from single color images. Diverse training data for this task can be automatically generated from sequences of a moving robot arm without any object in its gripper by following the procedure explained in Section 3.1. To prevent over-fitting on the background, we paste the extracted robot arm masks at random scale and horizontal translation on images of the MS COCO dataset [27]. Unlike [25] we do not incorporate backgrounds from the test environment into our training data. Furthermore, to emulate the presence of a grasped object we paste one random object crop (different from our test objects) per training sample at the gripper’s position. We derive these spots by opening or closing the gripper while the

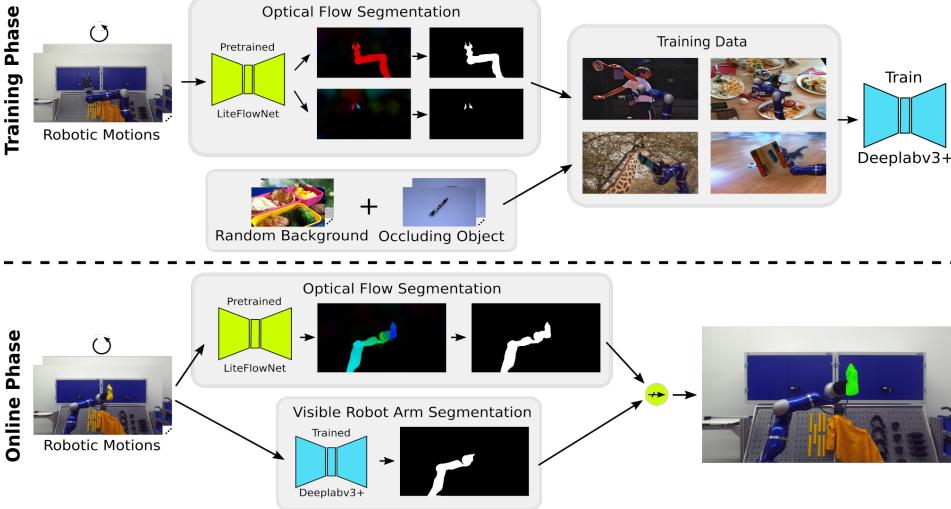


Figure 2: In a training phase we predict optical flow for robot arm motions using a pre-trained LiteFlowNet. By thresholding pixel-wise velocity we generate robot masks that are used to train a DeepLabv3+ to segment the visible robot arm from randomized backgrounds. We emulate the presence of an occluding object at the gripper’s position (top). The trained CNN is then used to separate the robot manipulator from any unknown, grasped object in the joint optical flow (bottom), where  $\rightarrow$  denotes a nimply gate. The result is an abundant source of labeled data that can be produced online in any static environment.

robot remains at a certain joint position and obtain a corresponding mask again by thresholding the flow magnitude - a procedure which is not necessary during inference. Whenever this gripper mask consists of two separated parts, we can further realistically emulate the grasp by placing the object above the smaller part (most likely further away from the camera, thus occluded) and overlay the item with the larger one (potentially closer to the camera, thus occluding the object). In all other cases we sample a random point on the gripper mask which denotes the object center. The occluding objects are not learned explicitly but belong to the background class. Our validation set consists of a disjoint set of robot poses and occluding objects. As we want to verify the generalization capability of our approach onto novel environments, we do not replace the background for these images. We generate 45,000 training and 5,000 validation images and train a DeepLabv3+ [28] on semantic segmentation. Afterwards, we are able to segment the visible robot arm regardless of the object in the gripper. Information on the occluder objects and additional training details are given in Section D.

### 3.3 Grasped Object Segmentation

During run time, we use the same flow network to derive a joint robot-object mask. Concurrently, we predict the visible robot arm with the trained Deeplabv3+. Finally, parts in motion but not detected as robot arm are denoted as the unknown object. We further refine object masks to encounter false-positive and false-negative predictions of both networks. First, we delete all parts in contact with the image border. We expect the object to be fully visible in the camera frame, and we hypothesize that our trained network successfully segments the area around the gripper. Hence, the object mask should be detached from all irrelevant mask fractions identified by optical flow thresholding. Second, we once more utilize the gripper masks which were derived by opening and closing the gripper at stationary robot arm positions during the visible robot arm segmentation. These spots are now being used as estimation of the object’s position: We always keep the part of the mask that is closest to this location, but reject every component that is further away than 100 pixels. Finally, we delete mask parts whose area is below 2500 pixels to remove small noisy artifacts.

### 3.4 Trajectory Generation

The main goal of our method is to automatically generate data for training object segmentation networks and for performing 3D object reconstruction. For both tasks it is important to obtain as

diverse object views as possible. A trajectory obtained e.g. by kinesthetic teaching could instead create object views biased towards specific poses.

Even though we do not assume that the object-camera transformation is known, we can still produce nearly equally distributed end-effector rotations using Fibonacci sampling [29] with  $N = 301$  2D points on a sphere (mathematical details are explained in Section C). From these points we can construct corresponding rotation matrices  $R_i$  via up and forward vectors. The robot would heavily occlude the object in some of the poses where the gripper points away from the camera. Therefore, we split the view sphere in half and mirror the gripper rotations pointing away from the camera to the side that points approximately towards the camera. To obtain views from all sides, the motion is repeated with the gripper rotated 180 degrees.

Additionally, some rotational end-effector movements do not create sufficient motion to reliably detect the object mask by means of optical flow. Yet, motion between two consecutive frames is of severe importance for successful segmentation. To alleviate this dependency one could propagate optical flow over multiple frames (e.g. [30]). However, this can lead to drift effects and propagation of erroneous regions, which potentially worsens an initial prediction. Therefore, we propose to continuously follow Cartesian points on an ellipse-shaped trajectory that is approximately parallel to the image plane. We repeatedly loop through  $n_e = 20$  equidistant points on the ellipse and use them as end-effector translations. Finally, we assign end-effector rotation matrices  $R_i$  by spiraling through the points on the sphere from the Fibonacci sampling.

## 4 Setup and Dataset Creation

The main setup consists of a KUKA LBR4+ robot arm on a linear axis with a Robotiq 2F-85 two-finger gripper. Images are recorded with a ZED stereo sensor.

### 4.1 Robotic Arm Recordings under Weak Camera Calibration

A major advantage of our method is the independence of a precise robot-camera calibration. Yet, we expect the robot to move inside the camera frame on an ellipse-shaped trajectory which should be approximately parallel to the image plane (Section 3.4). Furthermore, all parts of the object should be visible inside the frame’s borders to not be discarded during post-processing (Section 3.3). Nevertheless, these constraints can be satisfied quite effortless with approximate knowledge about the robot’s motion in the world coordinate frame, and by placing the camera loosely pointing towards the robot. We use the trajectory as described in Section 3.4 to record images of the robot at 301 poses. At every configuration  $P_i$  we shift the linear axis with a small offset to ensure that all parts of the robot jointly move at all times. Additionally, at every such pose we solely move the gripper jaws to different positions to determine spots where to paste occluding objects, as explained in Section 3.2.

### 4.2 Segmenting YCB-Video Objects

To demonstrate the effectiveness of our suggested method we record 15 different objects of the YCB-Video Dataset [31] while being grasped by a robotic gripper. Grasping of unknown objects is a challenging task on its own and often requires priors in the form of object poses [32]. As this is not the focus of our work we directly hand over the respective objects to the robot which could also be a realistic application scenario. Nevertheless, to further automate our approach we carry out a grasping study on these objects and refer the reader to Section B of the supplementary material for results. Note that we use the same trajectory as in Section 4.1 for a direct comparison of our method to change detection algorithms.

### 4.3 Baselines: Change Detection and Object-Agnostic Segmentation

As a baseline measure we explore pixel-wise change detection [7, 33] in RGB space ( $CD_{RGB}$ ) and between optical flow masks ( $CD_{OF}$ ), and direct object-agnostic segmentation (Obj-Agn). For change detection we record images of the robot arm both with and without an object at the same position and select pixels based on the absolute difference between them. In the RGB case we empirically set  $\tau = \frac{p}{25}$  with  $p$  denoting the pixel range; for optical flow masks this reduces to a binary selection. Regarding direct object-agnostic segmentation we train another DeepLabv3+ on semantic

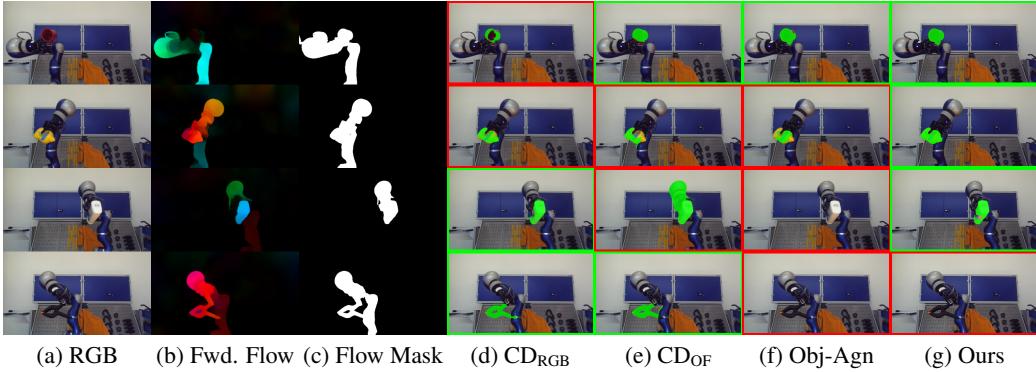


Figure 3: Qualitative results of our method and the three baselines (best viewed magnified and in color). Red borders denote poor performance, while green ones indicate satisfactory segmentation masks. Top row:  $CD_{RGB}$  tends to perform poorly whenever object and background are of similar color. Note that while  $CD_{OF}$  and Obj-Agn deliver decent results despite small cut-off parts, our method even successfully segments the handle (best viewed zoomed-in). Second row:  $CD_{OF}$  can only segment moving parts as object which are directly located in front of the background. Third row:  $CD_{OF}$  over-segmentation due to faulty LiteFlowNet predictions, which our method corrects. Fourth row: A failure case of our method where parts of the object are identified as robot due to the similar color and texture of clamp and gripper. Our applied post-processing deletes the remaining small blobs and results in an empty mask. Obj-Agn fails to segment the last two objects completely.

segmentation (similar to the one trained on robot arm segmentation), and additionally predict the object in the gripper as separate semantic class. During testing we can directly infer the binary mask of the unknown object from one single RGB frame. For fair comparison we apply the same post-processing on every method except for change detection in RGB space, where we perform morphological opening and closing.

## 5 Evaluation of Automatically Annotated Object Masks

We proceed to evaluate our automatic annotation method. For all segmentation results we report the standard *mean Intersection over Union* (mIoU). Therefore, we manually annotate 15% of all recorded YCB object images with ground truth segmentation masks. The object poses of this subset are identical across all items.

### 5.1 Comparison to Baselines

Table 1 reports the mIoU of our proposed method compared to three baselines. We outperform these by a considerable margin across all objects except one, which is best explained in Figure 3. The performance gain can be traced back to the fact that optical flow and robot segmentation networks are trained to extract features that are robust against noise, similar colors and non-rigid parts. We also circumvent the hard, potentially ill-defined task of learning how to segment unknown objects - here, an object is simply defined as a physically connected entity in motion. For further results on deformable objects like towels and detailed evaluation of our post-processing we refer the reader to Section A.1 of the supplementary material.

### 5.2 Comparison to Literature

To the best of our knowledge, the most similar approach in literature stems from Florence et al. [25]. Yet, they require robot-camera calibration as well as depth information and obtain their foreground mask with GrabCut, while the distinction between robot and object is done in a similar fashion with a CNN. We compare our method on a joint subset of YCB objects (see Table 1). Although we refrain from camera calibration and solely utilize a single RGB image stream, we are able to surpass their respective baseline.

Table 1: Quantitative comparison of our method compared to three baseline approaches and [25]; for a better overview we first list the joint objects. The two right-most columns show results on semantic segmentation and detection in a table top scene, where we use our automatically generated object masks as training data. Note that we compute the IoU for detection for a comparison to the semantic segmentation results - this is possible since no instance appears more than once per image. Numbers in **bold** denote the best results on object-in-gripper segmentation.

YCB Object	Object-in-Gripper Segmentation					<i>Ours</i> as training data	
	CD <sub>RGB</sub>	CD <sub>OF</sub>	Obj-Agn	[25]	Ours	Sem. Seg.	Detection
011_banana	33.72	47.13	00.00	54.60	<b>77.23</b>	75.18	54.63
019_pitcher_base	42.05	74.66	48.87	<b>91.90</b>	78.56	88.75	88.93
024_bowl	41.09	78.18	57.07	90.10	<b>91.19</b>	53.69	75.99
025_mug	39.38	77.26	84.73	84.30	<b>86.36</b>	52.68	86.22
035_power_drill	18.16	50.77	52.32	<b>62.60</b>	60.69	77.72	79.10
037_scissors	06.43	10.35	00.00	<b>62.70</b>	46.01	42.35	26.27
061_foam_brick	22.10	61.69	<b>87.29</b>	57.40	80.08	74.83	60.56
Average (joint objects)	28.99	57.15	42.91	71.94	<b>74.30</b>	66.46	67.39
003_cracker_box	34.06	70.60	57.50	-	<b>88.84</b>	74.29	75.19
005_tomato_soup_can	27.57	38.66	76.57	-	<b>81.53</b>	62.49	91.70
006_mustard_bottle	32.31	66.94	55.56	-	<b>86.24</b>	79.77	80.92
007_tuna_fish_can	17.87	52.69	00.00	-	<b>60.81</b>	51.27	42.99
008_pudding_box	46.10	64.89	76.66	-	<b>80.40</b>	73.78	78.64
010_potted_meat_can	34.81	71.07	75.46	-	<b>81.78</b>	63.87	82.53
021_bleach_cleanser	32.99	61.19	33.99	-	<b>83.82</b>	54.01	83.79
052_extra_large.clamp	38.98	52.37	00.00	-	<b>53.48</b>	58.68	46.45
Average (total)	31.18	58.56	47.07	-	<b>75.80</b>	65.56	70.26

### 5.3 Object Segmentation from Scenes

We use the generated object masks to train DeepLabv3+ on semantic segmentation from RGB scenes. We test our model on 100 images where we place the respective objects at random position and orientation on a table. Results are depicted in Figure 4 and in the second from right column in Table 1. Even though the test set was recorded in a different environment than the training data, the mIoU values in the test scene are approaching those of the generated in-gripper training data. The *pitcher* and *drill* even receive higher mIoU scores than when segmented in the gripper. We ascribe this to the ability of neural networks to average out annotation errors. However, although the *mug* and *bowl* stood out during the automatic mask generation, the network seems to have difficulties differentiating between them since they have very similar texture. We believe that this is because of incomplete masks which provide unreliable information on the object’s shape.

We also train MaskRCNN [34] on instance segmentation using the same automatically annotated training data (right-most column of Table 1) and qualitatively compare results to a MaskRCNN trained on the YCB Video dataset in Figure 4. It shows that our grasped object data generalizes to novel settings while training on YCB Video does not generalize to a novel sensor.

CNNs typically require many different views of an object to derive a good representation that generalizes well to novel scenes. To this end, we investigate the performance loss onto the same test scene given less than the initial  $n = 301$  almost equidistant views. While merely two views of an object provide sufficient information to achieve 60% of the final mIoU with 301 views, the results underline that view diversity indeed increases segmentation performance. For more results please see Section A.2 of the supplementary material.

### 5.4 Photogrammetric 3D Object Reconstruction

Textured 3D meshes contain rich information about objects. In robotics they can be utilized for grasp planning, object pose estimation or tracking and learning in simulation. A robot that can perform 3D object reconstruction without human intervention therefore strongly increases autonomy.

Photogrammetric 3D object reconstruction is usually performed by taking photos of a static object from various camera positions, finding correspondences, computing their 3D coordinates and thereof



Figure 4: Semantic / instance segmentation results. Top row: DeepLabv3+ trained on our grasped objects on semantic segmentation; Middle row: MaskRCNN trained on our grasped objects; Bottom row: MaskRCNN trained on the YCB Video dataset [31]. Networks trained on our acquired data usually generalize well to table top scenes with few failure cases due to thin parts or over-segmentation (first two rows). Networks trained on the original annotated YCB data, which consists of table top scenes recorded with a different sensor, fail to generalize due to the domain gap (bottom row).

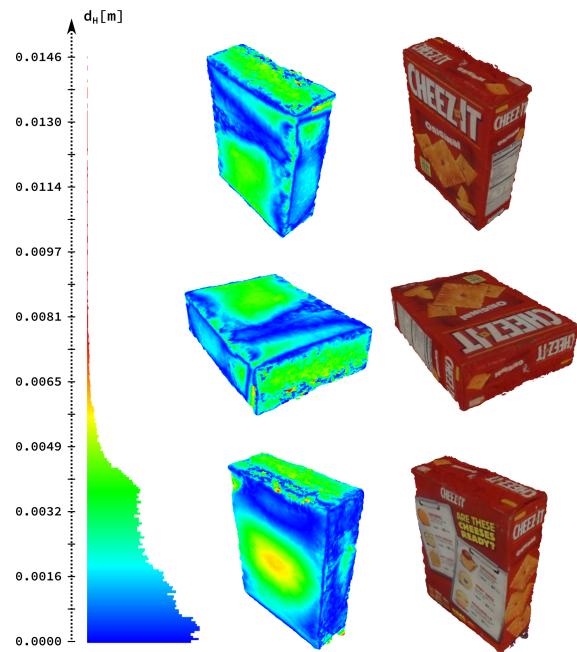


Figure 5: Photogrammetric 3D Reconstruction of *003\_cracker\_box* directly from the robotic gripper; Left: Histogram of Euclidean distances  $d_H$  from nearest vertices between ground truth mesh (YCB Video dataset [31]) and our aligned 3D reconstruction. The mean distance between vertices is  $\mu(d) = 0.2\text{cm}$ ; Middle: Quality Mapping; Right: Texture Mapping; Note that any potential gripper parts were successfully removed from the object.

building a mesh. Disadvantages of this method are that the bottom of the object is usually not captured and that the environment and object should be static at all times. Since our presented methods yield accurate segmentation masks, we can reverse the problem by actively recording views of a rotating object using a static camera. One challenge is that the lighting on the object is dynamic during the recording process. Nonetheless, our masked object views can be successfully fused without manual intervention or tuning using a commercial reconstruction software [35]. Figure 5 shows the resulting textured mesh of the YCB Cracker Box and the Euclidean distances of the nearest vertices on the aligned ground truth model. Please note that pure monocular photogrammetric 3D object reconstruction can also fail for some symmetric and texture-less objects. In these cases, information from depth cameras has to be incorporated.

## 6 Conclusion

We have presented a framework to automatically generate unseen object and robot arm segmentation masks through robotic interactions. Our optical flow based approach requires little prior knowledge and poses few constraints on the setup. Particularly, the camera does not need to be calibrated against the robot and thus can be freely moved. The usability of the generated labels has been demonstrated in several downstream tasks like semantic and instance segmentation from scenes and 3D object reconstruction. The latter is a step forward to close the real2sim2real cycle. Our work is motivated by both the way humans learn through object interactions and practical limitations of supervised learning algorithms that stem from the lack of annotated data. We conclude that observing self-generated motions can enable robots to semi-autonomously learn about newly encountered objects, which is a prerequisite for self-improving systems.

## Acknowledgments

We thank the reviewers for their useful comments. This work was partially supported by the DLR-internal project "Factory of the Future".

## References

- [1] T. Hodan, V. Vineet, R. Gal, E. Shalev, J. Hanzelka, T. Connell, P. Urbina, S. N. Sinha, and B. Guenter. **Photorealistic Image Synthesis for Object Instance Detection**. *arXiv preprint arXiv:1902.03334*, 2019.
- [2] M. Sundermeyer, Z.-C. Marton, M. Durner, M. Brucker, and R. Triebel. **Implicit 3d orientation learning for 6d object detection from rgb images**. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 699–715, 2018.
- [3] M. A. Lobo, E. Kokkoni, A. C. de Campos, and J. C. Galloway. **Not just playing around: infants' behaviors with objects reflect ability, constraints, and object properties**. *Infant Behav Dev*, 37(3):334–351, Aug 2014.
- [4] K. Nakayama and J. M. Loomis. **Optical Velocity Patterns, Velocity-Sensitive Neurons, and Space Perception: A Hypothesis**. *Perception*, 3(1):63–80, 1974.
- [5] T.-W. Hui, X. Tang, and C. C. Loy. **LiteFlowNet: A Lightweight Convolutional Neural Network for Optical Flow Estimation**. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8981–8989, June 2018.
- [6] J. Bohg, K. Hausman, B. Sankaran, O. Brock, D. Kräig, S. Schaal, and G. S. Sukhatme. **Interactive Perception: Leveraging Action in Perception and Perception in Action**. *IEEE Transactions on Robotics*, 33(6):1273–1291, 2017. doi:10.1109/TRO.2017.2721939.
- [7] A. Singh. **Review Article Digital change detection techniques using remotely-sensed data**. *International Journal of Remote Sensing*, 10(6):989–1003, 1989.
- [8] J. Kenney, T. Buckley, and O. Brock. **Interactive Segmentation for Manipulation in Unstructured Environments**. In *Proceedings of the 2009 IEEE International Conference on Robotics and Automation, ICRA'09*, pages 1343–1348. IEEE Press, 2009.
- [9] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis. **Background and foreground modeling using nonparametric kernel density estimation for visual surveillance**. *Proceedings of the IEEE*, 90(7):1151–1163, July 2002.
- [10] T. Bouwmans, S. Javed, M. Sultana, and S. K. Jung. **Deep Neural Network Concepts for Background Subtraction: A Systematic Review and Comparative Evaluation**. *Neural Networks*, 117:8–66, 2019.
- [11] Jianbo Shi and J. Malik. **Motion segmentation and tracking using normalized cuts**. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1154–1160, Jan 1998.
- [12] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1):185–203, 1981.
- [13] Z. Tu, W. Xie, D. Zhang, R. Poppe, R. C. Veltkamp, B. Li, and J. Yuan. **A survey of variational and cnn-based optical flow techniques**. *Signal Processing: Image Communication*, 72:9–24, 2019.
- [14] K. Fragkiadaki, P. Arbelaez, P. Felsen, and J. Malik. **Learning to segment moving objects in videos**. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4083–4090, 2015.
- [15] P. Bideau, A. RoyChowdhury, R. R. Menon, and E. Learned-Miller. **The Best of Both Worlds: Combining CNNs and Geometric Constraints for Hierarchical Motion Segmentation**. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

- [16] A. Dave, P. Tokmakov, and D. Ramanan. *Towards Segmenting Everything That Moves*. *CoRR*, abs/1902.03715, 2019.
- [17] C. Xie, Y. Xiang, D. Fox, and Z. Harchaoui. *Object Discovery in Videos as Foreground Motion Clustering*. *CoRR*, abs/1812.02772, 2018.
- [18] L. Shao, P. Shah, V. Dwaracherla, and J. Bohg. *Motion-based Object Segmentation based on Dense RGB-D Scene Flow*. *CoRR*, abs/1804.05195, 2018.
- [19] P. Fitzpatrick and G. Metta. *Grounding vision through experimental manipulation*. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 361(1811):2165–2185, Aug. 2003.
- [20] D. Schiebener, A. Ude, and T. Asfour. *Physical interaction for segmentation of unknown textured and non-textured rigid objects*. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4959–4966, May 2014.
- [21] D. Schiebener, A. Ude, J. Morimoto, T. Asfour, and R. Dillmann. *Segmentation and learning of unknown objects through physical interaction*. In *2011 11th IEEE-RAS International Conference on Humanoid Robots*, pages 500–506, Oct 2011.
- [22] D. Pathak, Y. Shentu, D. Chen, P. Agrawal, T. Darrell, S. Levine, and J. Malik. *Learning Instance Segmentation by Interaction*. In *CVPR Workshop on Benchmarks for Deep Learning in Robotic Vision*, 2018.
- [23] G. Du, K. Wang, and S. Lian. *Vision-based Robotic Grasping from Object Localization, Pose Estimation, Grasp Detection to Motion Planning: A Review*. *CoRR*, abs/1905.06658, 2019.
- [24] C. C. d. Rocha, N. Padoy, and B. Rosa. *Self-Supervised Surgical Tool Segmentation using Kinematic Information*. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8720–8726, 2019.
- [25] V. Florence, J. J. Corso, and B. Griffin. *Self-Supervised Robot In-hand Object Learning*. *CoRR*, abs/1904.00952, 2019.
- [26] N. Otsu. *A Threshold Selection Method from Gray-Level Histograms*. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, Jan 1979.
- [27] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. *Microsoft COCO: Common Objects in Context*. *CoRR*, abs/1405.0312, 2014.
- [28] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. *Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation*, 2018.
- [29] Á. González. *Measurement of Areas on a Sphere Using Fibonacci and Latitude–Longitude Lattices*. *Mathematical Geosciences*, 42(1):49–64, Nov 2009.
- [30] T. K. Mai, M. Gouiffès, and S. Bouchafa. *Optical Flow Refinement using Reliable Flow Propagation*. In *VISIGRAPP*, 2017.
- [31] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. *PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes*. *CoRR*, abs/1711.00199, 2017.
- [32] J. Shi and G. S. Koonjul. *Real-time Grasp Planning with Environment Constraints*. In *IROS 2014 Workshop: Real-time Motion Generation & Control*, 2014.
- [33] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam. *Image change detection algorithms: a systematic survey*. *IEEE Transactions on Image Processing*, 14(3):294–307, 2005.
- [34] K. He, G. Gkioxari, P. Dollár, and R. Girshick. *Mask R-CNN*. *arXiv preprint arXiv:1703.06870*, 2017.
- [35] L. Agisoft. *Agisoft metashape user manual, Professional edition, Version 1.5*. Agisoft LLC, St. Petersburg, Russia, accessed June, 2018.
- [36] S. Maji. *Large Scale Image Annotations on Amazon Mechanical Turk*. Technical Report UCB/EECS-2011-79, EECS Department, University of California, Berkeley, Jul 2011.

## A Additional Results

### A.1 Deformable Objects

Our proposed method does not require strict rigid objects. To demonstrate this we record four deformable objects (Figure 6) on half of the points in the proposed trajectory (i.e., one half side of the sphere in Figure 9). For results see Table 2 and Figure 7.



Figure 6: Four additional, non-rigid objects. From left to right, top to bottom: towel, mask, notebook, cable.

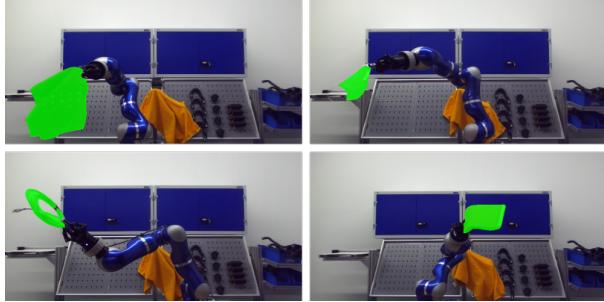


Figure 7: Exemplary quantitative results on the four deformable objects. From left to right, top to bottom: towel, mask, cable, notebook.

Table 2: Additional quantitative results on deformable objects. All values are in %.

Deformable Object	mIoU	Precision	Recall
cable	50.95	63.44	72.13
mask	69.79	72.10	95.63
notebook	77.24	93.41	81.69
towel	81.81	93.19	87.02
Average	69.95	80.53	84.12

### A.2 Number of Views for Segmentation Performance

It is expected that many distinct views are beneficial for the generalization ability of neural networks. We explore this by training a network with our automatically segmented objects on semantic segmentation in a table-top setting, and increment the amount of object views across different trainings. Figure 8 shows the respective results.

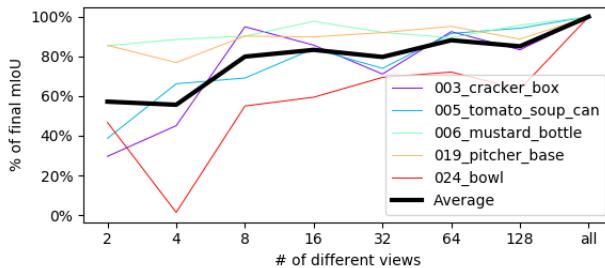


Figure 8: Number of different views in relation to the relative final mIoU with all object views for five different objects. With only two viewpoints the network achieves on average 60% of the final mIoU (301 poses), and there is a notable correlation between performance and number of views.

## B Grasping of Unknown Objects

To further automate our segmentation pipeline we evaluate grasping and re-grasping after placing success on the YCB objects in Table 3. The re-grasping at  $180^\circ$  degree gripper rotation is necessary as we want to record the backside of the object. Here, solely the object’s position on the table is known to the robot. Only the *bowl* cannot be grasped at a static position from top-down. The *scissors* is unstable and slightly offset when placed back onto the table so that the re-grasp at  $180^\circ$  gripper rotation can fail. Apart from these issues, the results confirm that our strategy is successful for almost all objects and depicts a practical way to minimize human intervention.

Table 3: Top-Down grasp successes for YCB objects. *Re-graspable* refers to placing the object and re-grasping it at  $180^\circ$  rotated gripper position.

YCB Object	graspable	re-graspable
003_cracker_box	✓	✓
005_tomato_soup_can	✓	✓
006_mustard_bottle	✓	✓
007_tuna_fish_can	✓	✓
008_pudding_box	✓	✓
010_potted_meat_can	✓	✓
011_banana	✓	✓
019_pitcher_base	✓	✓
021_bleach_cleanser	✓	✓
024_bowl	✗	✗
025_cup	✓	✓
035_power_drill	✓	✓
037_scissors	✓	✗
052_extra_large_clamp	✓	✓
061_foam_brick	✓	✓
Total	14/15	13/15

## C Fibonacci Sampling

First we create an evenly distributed lattice using  $N = 301$  2D points with coordinates

$$(x_i, y_i) = \left( \frac{i + 1/2}{N}, \frac{i}{\phi} \right) \quad \text{for } 0 \leq i \leq N - 1 \quad (2)$$

where  $\phi = \frac{1+\sqrt{5}}{2}$  is the golden ratio. Then, we perform an area preserving mapping onto a cylinder and subsequently onto a sphere by

$$(x, y) \rightarrow (\theta, \phi) : \quad (\cos^{-1}(2x - 1) - \pi/2, 2\pi y) \quad (1)$$

$$(\theta, \phi) \rightarrow (x, y, z) : \quad (\cos \theta \cos \phi, \cos \theta \sin \phi, \sin \theta) \quad (2)$$

to obtain the respective coordinates on the sphere. We construct corresponding rotation matrices  $R_i$  via up and forward vectors. A visualization of a reduced set of almost equidistant points on a sphere is depicted in Figure 9.

### C.1 Ablation Study of our Post-Processing

In Table 4 we investigate the influence of different design choices for our method regarding data generation / training setting and post-processing on the YCB objects.

The main success factors for our final metrics are the pasting of a distractor object, the maximum gripper distance and the minimum mask size. There is, albeit small, a notable difference between

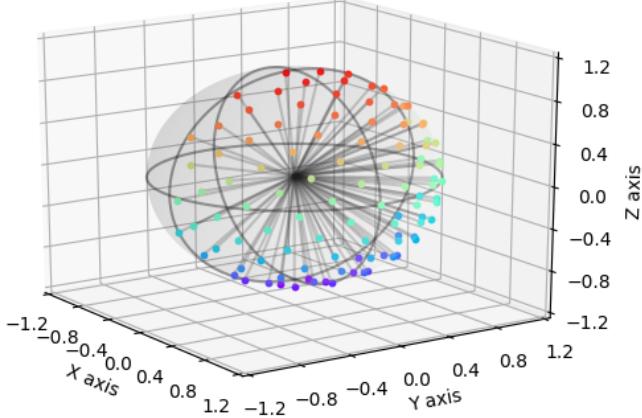


Figure 9: Visualization of Fibonacci sampling of almost equidistant points on a sphere. We only consider the end-effector rotations pointing approximately towards the camera and then regrasp the object for the back side of the sphere.

Table 4: Ablation study on different data augmentation and training settings (*D/T*) and post-processing (*PP*) across different flow masks (from which we subtract the robot prediction to derive an object mask). In every consecutive row we drop the respective step and all previous ones. Altogether we are able to boost object-in-gripper segmentation from an initial object mIoU of around 51% by about 25 percent points. Numbers in **bold** denote the best results.

		Forward Flow	Intersection	Union
	Ours	<b>76.32</b>	72.15	75.80
PP	- Min. mask size	65.54	57.29	<b>68.04</b>
	- Max. grip. dist.	60.01	53.63	<b>61.50</b>
	- Border deletion	57.61	53.47	<b>58.29</b>
D/T	- Gripper loss weight	56.24	51.22	<b>57.55</b>
	- Occluding object	<b>51.79</b>	48.14	51.70

forward flow and forward/backward union, and at its best, forward flow beats cumulative masks by a small margin. However, this is merely a result of discarding information in the last post-processing step and we therefore used the cumulated masks. Flow intersection masks between forward and backward flow, on the other hand, might eliminate noise but heavily rely on motion at the same spatial positions between consecutive frames. Even with perfect continuous movement this can not be guaranteed due to potential mis-predictions of the LiteFlowNet.

## C.2 mIoU, Precision and Recall between Vanilla and Post-Processed Results

Table 5 lists the gain in % for different metrics for three baselines and our proposed approach.

## D Training Details

As occluding objects we use objects from the BigBird dataset (<http://rll.berkeley.edu/bigbird/>) (discarding instances of the YCB Video dataset), and a set of internally recorded objects depicted in Figure 10. During training we apply random color jitter on the occluding objects with focus on

Table 5: Quantitative comparison between vanilla results and applied post-processing. The metrics denote the respective average across all object classes. Especially noteworthy are the small differences between our vanilla and CD<sub>OF</sub> post-processed results. All values are in %. Numbers in **bold** denote the best results.

Method	Vanilla			With Post-Processing		
	mIoU	Precision	Recall	mIoU	Precision	Recall
CD <sub>RGB</sub>	10.88	11.57	72.66	31.18	60.19	39.31
CD <sub>OF</sub>	38.71	44.87	73.84	58.56	69.53	78.35
Obj-Agn	34.36	<b>72.22</b>	38.24	47.07	66.94	50.27
Ours	<b>58.29</b>	63.79	<b>85.42</b>	<b>75.80</b>	<b>81.94</b>	<b>90.60</b>



Figure 10: Occluding objects used in our training data.

blueish augmentation to account for over-segmentation on our blue-biased robot arm. To ensure that the model is able to differentiate well between gripper and object, we weight the area around the gripper with a Gaussian shape when calculating the loss. The respective center is again obtained from the gripper spots. We empirically set the width to 50 pixels and the magnitude at the center to a factor of 3.

We use a DeepLabv3+ implementation in PyTorch pre-trained on ImageNet. For all object-in-gripper segmentation tasks we use the default learning rate with a polynomial decay and Adam optimizer. We train with a fixed input size of 414x736 pixels.

For all optical flow predictions we use a LiteFlowNet implementation ([Link to GitHub](#)) pre-trained on Things3D and Chairs. LiteFlowNet is our flow predictor of choice due to its superior performance on popular benchmark datasets at time of release and fast flow estimation. For an exhaustive comparison between conventional, hybrid and network-based optical flow estimators we refer the reader to [5], Table 2.

For semantic segmentation we use the same DeepLabv3+ but freeze the encoder. The decoder is trained with a reduced learning rate of 0.0007. Similar to the robot arm segmentation, we paste multiple instances on MS COCO images at random scale, translation and in-plane rotation and adapt the labels accordingly. We allow occlusions up to 60% of each object instance mask and re-sample the training image if this limit is exceeded. Thereby, we produce 50,000 images. As the objects differ in size, we weight the object classes inversely proportional by the number of pixels to put equal focus on smaller objects, and clip the background weight to the factor of the largest object class.

We train the MaskRCNN with the standard settings and refer the reader to the original paper [34] for further details.

## E Estimated Setup and Inference Time

Regarding the aspect of setup time vs. manual annotations, users spent on average 61 seconds to draw the outline of a single object for the PASCAL VOC 2010 trainval dataset [36]. Assuming similar timings this would result in more than 76 hours if we would manually label every image of our recorded YCB objects (301 images / 15 objects). Due to the fact that our algorithm needs no precise calibration, we expect implementing the setup to only take a few hours (excluding training time). Then, a human merely needs to place objects at marked positions for top-down grasping, or hand the respective object to the robot, which should take no longer than a minute per object. Afterwards, there is no additional human involvement and the camera can be moved. We conclude that while there is a qualitative difference between manual and our self-supervised annotations, our method is to be preferred regarding setup time.

In inference mode, our method runs at approximately 5 fps on a single Nvidia GeForce RTX 2080 Ti.

## F YCB Objects used in this work

Figure 11 shows all 15 YCB Video objects used in this work.



Figure 11: The 15 items from the YCB Video Objects which are used in our experiments.