

“What’s This?” - Learning to Segment Unknown Objects from Manipulation Sequences

Wout Boerdijk¹

Martin Sundermeyer¹

Maximilian Durner¹

Rudolph Triebel^{1,2}

Abstract— We present a novel framework for self-supervised grasped object segmentation with a robotic manipulator. Our method successively learns an agnostic foreground segmentation followed by a distinction between manipulator and object solely by observing the motion between consecutive RGB frames. In contrast to previous approaches, we propose a single, end-to-end trainable architecture which jointly incorporates motion cues and semantic knowledge. Furthermore, while the motion of the manipulator and the object are substantial cues for our algorithm, we present means to robustly deal with distraction objects moving in the background, as well as with completely static scenes. Our method neither depends on any visual registration of a kinematic robot or 3D object models, nor on precise hand-eye calibration or any additional sensor data. By extensive experimental evaluation we demonstrate the superiority of our framework and provide detailed insights on its capability of dealing with the aforementioned extreme cases of motion. We also show that training a semantic segmentation network with the automatically labeled data achieves results on par with manually annotated training data. Code and pretrained model are available at <https://github.com/DLR-RM/DistinctNet>.

I. INTRODUCTION

One of the key functionalities for any robotic manipulation system is the ability to accurately determine the location of the object(s) to be manipulated. To do this, most current systems employ computer vision techniques, where the aim is to find the pixels that correspond to the target object within a camera image taken from the scene. This is called *object segmentation*, and it is an important prerequisite for a large number of downstream tasks such as object pose estimation, 3D reconstruction, or grasp detection. In general, object segmentation can be treated in two different ways, depending on whether a representation of the object is given beforehand, for example in form of a 3D model. If such a model exists, the segmentation problem is significantly easier, because the model can be used to guide the search, e.g. by training an object detector. Unfortunately, it is also the less common case, because 3D models are often not available.

Therefore, in this work we address segmentation of unknown objects, where the only cue we rely on is the change of pixels that occurs when moving the object in front of the camera. This can either be done by pushing the object on a table [1, 2, 3, 4] or by holding it with a gripper and moving it with the manipulator. We follow the latter approach, because then the object can be seen from all sides, including from the bottom. We also employ *self-supervised* object segmentation,

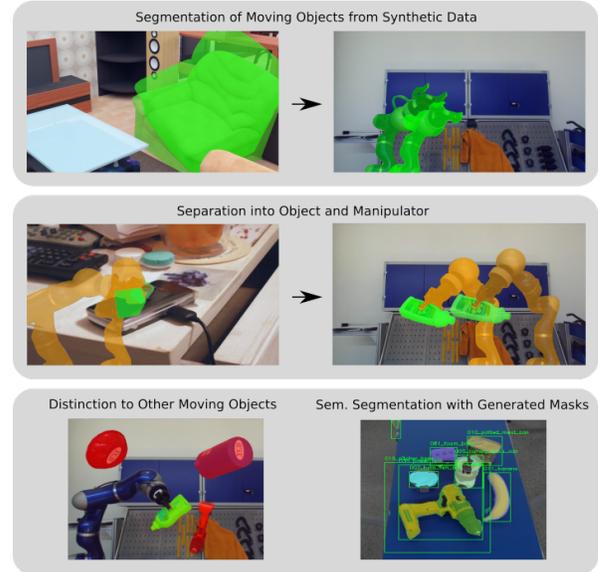


Fig. 1: Our approach first learns to segment objects that move in the foreground (top). Then, we use fine-tuning to learn a segmentation between the manipulator and the unknown object (center). Furthermore, our model can distinguish moving background objects (bottom left) and support downstream tasks such as semantic segmentation (bottom right).

i.e. for segmentation we neither require a kinematic model of the manipulator nor an intrinsic or extrinsic calibration of the camera. This makes our approach particularly versatile and easy to use. Also, in contrast to existing works, we do not perform a two-step process where first a representation of the manipulator is learned, before the actual object segmentation can be done. Instead, we propose an end-to-end learning scheme where the segmentation of both the manipulator and the object are learned simultaneously. This has the major advantage that information can be shared between the two tasks, and hyperparameter tuning or post processing are not required. Furthermore, our approach is able to distinguish between objects that are actively moved by the manipulator and those that move in the background (see Fig. 1). Finally, our architecture provides means to segment the object even in cases where no motion is observed for a small number of frames. As we will show experimentally, our approach outperforms previous self-supervised methods, and it even competes with techniques that rely on hand-eye calibration and additional depth data.

¹Institute of Robotics and Mechatronics, German Aerospace Center (DLR), 82234 Wessling, Germany <first>.<second>@dlr.de

²Department of Computer Science, Technical University of Munich (TUM), 85748 Garching, Germany

II. RELATED WORK

We briefly discuss existing work on self-supervised object segmentation. Additionally, we review methods that reason about correspondences between a pair of images. Relevant to this work are also approaches which combine motion and semantic segmentation.

A. Self-Supervised Grasped Object Segmentation

A robot arm has initially been utilized for object segmentation by inducing motion in a static scene (e.g. [1, 2, 3, 4]). To have more control over the object motion, grasping the respective item is a reasonable consequential step - however, it additionally requires distinguishing the object from the manipulator. This can be addressed with the help of a 3D robot model which allows to isolate the object [5, 6]. Browatzki et al. [7] employ a perception-driven object recognition framework where the robot itself controls the object views, and segment the item area with a GMM. Instead, [8] propose a self-supervised approach for pixel-wise robot recognition: By projecting a robot model onto the image plane and simultaneously optimizing a GrabCut-based cost function, segmentation labels of the robot arm are obtained. Florence et al. [9] build upon this framework and extend it to grasped object segmentation: They segment the foreground by projecting link positions of the robot into the camera frame which aids a graph-based depth segmentation followed by an additional refinement in RGB space. With this approach they collect segmentation masks of the manipulator to learn a robot arm representation. During inference they obtain a joint robot-object mask with the same foreground segmentation, and subtract the manipulator with help of the robot arm segmentation network. In our previous work [10] we alleviate the constraints of the aforementioned methods in form of precise camera calibration and additional sensor modalities by observing motion in a static scene, and learn a manipulator representation from thresholded optical flow only. Excluding the manipulator estimation from the joint object-manipulator mask - again derived by motion - yields the final object mask.

Yet, optical flow incorporates little semantic knowledge which requires additional post-processing. Moreover, none of the above approaches utilize a single, end-to-end trainable architecture.

B. Exploring Correspondences between Images

Finding dense correspondences between a pair of images is one of the fundamental tasks of computer vision. Input pairs can be different views of a scene [11, 12], consecutive video frames [13, 14] or even an image pair depicting different instances of the same semantic class as in the case of semantic correspondence matching [15, 16, 17]. We aim to segment moving objects from otherwise static background; hence we focus on the second category where change detection [18] or background modeling [19] are fundamental algorithms. Due to superior performance of neural networks in many image-related tasks, *Convolutional Neural Networks (CNNs)* have also been applied [20]. De Jong et al. [21] target the

problem of unsupervised change detection in satellite images and propose a siamese architecture with feature fusion by concatenation. Ru et al. [22] merge CNN features with a canonical correlation analysis to detect semantic changes in urban regions.

Optical flow is another way of obtaining pixel-wise correspondences between images, and current state-of-the-art approaches usually employ correlation in deep feature space to obtain a dense flow map (e.g. [23, 24, 14]).

Since the rise of attention [25] several works explore differences between two data features with so-called *co-attention*. This has initially been studied for vision-language tasks [26, 27], but already has been applied successfully to the image domain: Lu et al. [28] use co-attention for semantic video segmentation and achieve state-of-the-art performance for unsupervised video segmentation on DAVIS [29] at their time of release.

C. Combination of Motion and Segmentation

Many dense correspondence estimators have successfully been coupled with specific semantic knowledge. For instance, Cioppa et al. [30] propose a background modeling algorithm that is updated with predictions of a semantic segmentation network which has learned important object-related features and aids the overall pipeline. Similar work exist for optical flow estimation: [31] apply prior knowledge on a pixel's motion by defining different motion models for specific semantic regions. Bai et al. [32] segment potentially moving objects, estimate flow individually and merge the predictions with background motion to improve accuracy. In [33], classical geometric knowledge of moving objects is combined with high-level semantic image understanding in order to improve performance. Hur et al. [34] utilize flow to enforce temporal segmentation consistency, and simultaneously pose epipolar constraints inferred from semantic information on the motion estimation. Cheng et al. [35] propose to jointly predict optical flow and semantic segmentation masks by employing a two stream approach, while [36, 37] segment moving objects by employing a specific motion and appearance network.

Nevertheless, the current best performing methods on the Kitti Optical Flow Benchmark [38] solely estimate optical flow. Concurrently we observe that flow architectures usually differ from semantic segmentation networks, as features for each task are extracted differently. Moreover, we argue that optical flow itself over-complicates moving foreground extraction for object segmentation: for a respective binary mask it is solely interesting whether an object moves or not; direction and magnitude of each pixel do not yield any benefits.

Consequently, we rethink the utilization of motion in our self-supervised object segmentation pipeline. We propose an architecture that can efficiently estimate dense correspondences between features, and is additionally capable of incorporating specific semantic attributes. In contrast to previous self-supervised grasped object segmentation approaches our framework consists of a single network that is trainable end-to-end, and does not require any post-processing.

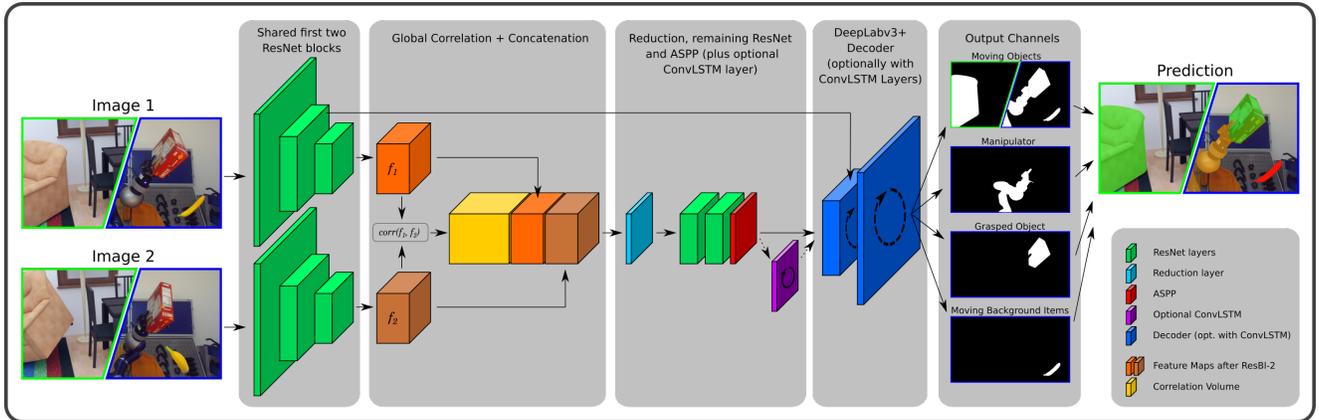


Fig. 2: Architectural overview (best viewed magnified). *DistinctNet* can be trained on moving foreground segmentation (left image cuts; green border) and afterwards fine-tuned on semantically distinguishing foreground into manipulator, object and optional other semantic classes such as moving background objects (right image cuts; blue border). Dashed lines denote optional recurrent layers. Note that the last three output channels (here depicting manipulator, object and a moving background object) are only available during fine-tuning.

III. METHOD

We modify a standard DeepLabv3+ [39] with a ResNet-50 [40] backbone which has proven to be successful for many semantic segmentation applications. In the following we refer to our network as *DistinctNet*, as it allows to (1) distinguish moving foreground from static background as well as (2) semantically differentiate objects. We now further explain our modifications that let us predict moving foreground and allow for fine-tuning on robot-object segmentation. For an overview of our pipeline please see Fig. 2.

A. Architecture

DistinctNet takes as input two images depicting the displacement of an object and otherwise static background. This input pair is processed by the first two ResNet blocks in siamese fashion with shared weights. A correlation layer then calculates the similarity between the two feature volumes to derive information on possible displacement, which is commonly done in disparity or optical flow estimation architectures (e.g. [41, 42]). Inspired by [16] we normalize features before and after the correlation operation, and finally apply a *ReLU* non-linearity. The output thereof is concatenated with the two feature maps and channel-wise reduced to fit the input dimension of the following ResNet block. Finally, forwarding through the remaining DeepLabv3+ modules generates pixel-wise probabilities of foreground and background. Additional output channels concatenated to the final layer of the network are utilized to predict different parts of the moving foreground like manipulator and object.

Furthermore, we experiment with *Convolutional LSTM* (*ConvLSTM*) layers [43] after the ASPP module and in the Decoder which let us predict object and manipulator even if their motion between consecutive frames comes to a halt. We investigate the performance of these recurrent layers in Sec. V-B.2. Last but not least, we evaluate the placement of

the correlation operation in our network in Sec. V-C, where we also consider employing a co-attention layer [26, 28].

B. Moving Object Segmentation

We first train *DistinctNet* to learn a semantic agnostic representation of a moving object in front of static background. While there exist several segmentation datasets with moving items in a scene [44, 29, 45], they are either limited to specific items, have multiple moving objects or additionally move the camera. Consequently, we use Blender-Proc [46] to generate realistic-looking synthetic image pairs of static scenes with one random object in motion between two consecutive frames, which we explain in full detail in Sec. IV-A.1. The synthetic image pairs let the network effectively reason about an unknown moving object in an otherwise static scene.

C. Distinction between Manipulator and Grasped Object

After training moving foreground estimation, *DistinctNet* is well suited to distinguish a moving manipulator without any grasped object from static background. Following the self-supervised paradigm these automatically generated masks are further used as training data, and we explain the respective details in Sec. IV-A.2. Importantly, the architecture of *DistinctNet* directly allows for fine-tuning on object-manipulator differentiation by increasing the number of output channels in the final layer. Now, the network is capable of moving foreground segmentation as well as semantically interpreting and distinguishing moving parts into manipulator and object in the additional channels.

The semantic foreground separation can be pushed even further: Remarkably, it allows us to resolve the constraint on static background as we can denote an additional channel for all moving items other than the grasped object and manipulator. Our aim here is to enhance the practical applicability of our approach, given that a completely static background

cannot always be assured. An experimental evaluation of this is given in Sec. V-B.1.

IV. DATASETS AND IMPLEMENTATION DETAILS

A. Training Datasets

1) *Moving Object Segmentation*: With BlenderProc [46] we sample a camera position inside a random room of the SunCG dataset [47], select an object inside the camera’s view and render ten frames while we randomly shift and rotate the respective item. We repeat this procedure to obtain 5,000 sequences of random moving objects in a static scene. DistinctNet is then trained on moving object segmentation (90/10 train/val split), where an input pair consists of two images of one sequence.

2) *Fine-Tuning on Robot-Object Segmentation*: The training data for this task should consist of moving robotic arms with a grasped object in front of random, stationary background. Hence, for each training pair we duplicate a random image of the MS COCO dataset [48] to emulate diverse static background. We further sample two random robot arms (one for each image) and augment the presence of an object at the gripper’s end effector with an occluding object. As in our previous work [10] we derive the corresponding object pasting spots by opening and closing the gripper at static position and once again segment moving foreground with DistinctNet. As occluders we use random objects from the ShapeNet dataset [49] whose orientation, texture and lighting can be realistically simulated with BlenderProc. Note that while we use the same object for both images in a training pair its orientation can be different. In total we generate 50,000 sample pairs (90/10 train/val split).

During training we separately augment each image with Gaussian noise, and add random color jitter for both the whole image as well as the occluding object. Additionally, to cope with pasting artifacts which can present a simple shortcut to the network during learning [50], we apply a median blur on the full image.

For our validation set we do not cut out robot masks, but directly paste objects on the test backgrounds. We do not augment these images to verify the generalization capability onto novel environments and objects.

B. Implementation Details

The global correlation layer restricts us to a fixed input size, and we consequently resize images to 414x746 pixels. All experiments are conducted on a single Nvidia GeForce RTX 2080 Ti. We use AdamW [51] as optimizer with a weight decay of 0.01. For moving object segmentation we use standard cross entropy loss and set the learning rate to 1e-3 with a reduction to 1e-4 for all ResNet layers. When fine-tuning on semantic classes we freeze all encoder weights before the correlation operation and train with a multi-class binary cross entropy loss combined with a reduced learning rate of 1e-4 for all trainable parameters. Training takes roughly four (moving object segmentation) and three (semantic finetuning) days with a batch size of 2 on a single Nvidia GeForce RTX 2080.

C. Recording Setup and Test Dataset

Our robot arm is a KUKA LBR4+ on a linear axis employed with a Robotiq 2F-85 two-finger gripper. All data is recorded with a ZED stereo sensor.

We use the same 15 YCB object recordings as in our previous work (301 images per object), and refer the reader to [10] for a detailed recording protocol. 45 images per object already contain manually annotated segmentation masks, and we additionally label the respective robot arms for a complete evaluation of foreground, robot and object.

V. EXPERIMENTAL EVALUATION

In this section we evaluate the performance of our method on self-supervised grasped object segmentation. We further provide insights on the model’s capability to deal with moving background objects and static input frames. We continue by exploring different architectural settings to explain our network design choices. Finally we show that the generated masks show equal or better performance for the task of semantic segmentation from scenes in comparison to a subset of ground truth annotations.

A. Evaluation of Self-Supervised Object Segmentation

We quantitatively evaluate our method in Tab. I and report the *mean Intersection over Union (mIoU)* for all objects as well as robot arm, moving foreground and background. Adding the gripper as separate semantic label (Sec. IV-A.2) increases overall object mIoU by one percent point.

DistinctNet generates strong segmentation masks (left part of Fig. 3), and we outperform recent self-supervised approaches with similar weak constraints on the setup by a large margin. Despite not using camera calibration, manipulator key point registration and additional depth data like [9] we still reach similar performance on a joint subset of objects (82.72% vs. 84.61% mIoU).

Nevertheless, some object views are quite challenging to our approach (middle part of Fig. 3), and a particularly outstanding item is the *037_scissors*. We hypothesize that due to the thin shape of the object the network finds it difficult to establish correspondences between two consecutive frames and identify it as moving object. This would also explain the superior performance of [9] on this item, as their approach does not rely on identification by motion.

We concurrently observe that fine-tuning on semantics improves moving foreground extraction (right part of Fig. 3): Initially, DistinctNet segments joint robot-object masks from our test set with a mIoU score of 86.56%. As reported in Tab. I this increases by roughly six percent points after fine-tuning, indicating that semantic knowledge (albeit here merely in the form of background, object and manipulator) is important for segmentation from motion, which is in line with previous research (Sec. II-C). Furthermore, while we hypothesize that the correlation operation itself ensures strong generalization from synthetic to real data, the semantic fine-tuning might also be helpful for bridging a remaining sim2real gap.

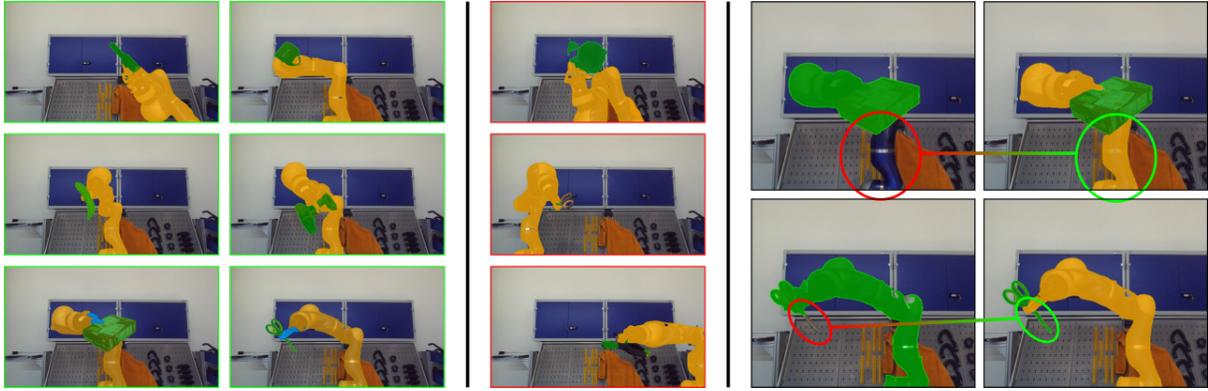


Fig. 3: DistinctNet generates semantic annotations of a moving manipulator with a grasped object and optional other classes like the gripper (left images); Some items are particularly challenging because of similar color to surroundings or due to their shape (middle images); Fine-tuning on semantics increases performance on moving foreground extraction (right images).

B. Extreme Cases of Motion

Our method typically relies on motion of manipulator and object. Nevertheless, small modifications to our framework allow to deal with challenging scenarios such as moving background items or no observable motion at all.

1) *Dealing with Non-Static Background Objects:* We can efficiently learn the presence of other moving objects by treating them as a new semantic class and adding another output channel to our network during semantic fine-tuning. We emulate random background object motion by pasting additional objects onto our training data with different random spatial positions between consecutive frames. After training, DistinctNet successfully differentiates a grasped object from other moving distractor items, as visualized in Fig. 4. Thereby, the overall object mIoU is merely reduced by 3% due to mismatches between the grasped item and an object moving in the background. This makes our approach

TABLE I: Mean IoU ([%]) of our method compared to change detection and two recent approaches from literature. *Ours+G.* denotes predictions with an additional gripper output channel. Numbers in **bold** depict best results.

Semantic Class	CD	[9]	[10]	Ours	Ours+G.
Object	33.24	-	75.80	83.11	84.21
003_cracker_box	62.58	-	88.84	90.57	92.70
005_tomato_soup_can	24.86	-	81.53	85.95	87.66
006_mustard_bottle	47.61	-	86.24	87.21	89.14
007_tuna_fish_can	13.95	-	60.81	75.44	74.89
008_pudding_box	32.34	-	80.40	90.28	91.74
010_potted_meat_can	26.18	-	81.78	85.39	84.68
011_banana	28.54	65.40	77.23	83.14	86.36
019_pitcher_base	32.84	95.70	78.56	89.47	89.43
021_bleach_cleanser	59.32	-	83.82	85.95	88.95
024_bowl	49.75	95.10	91.19	93.21	92.60
025_mug	33.16	93.10	86.36	91.93	92.36
035_power_drill	35.06	84.40	60.69	74.72	75.86
037_scissors	7.670	72.10	46.01	48.18	49.88
052_extra_large_clamp	23.35	-	53.48	72.88	74.35
061_foam_brick	21.33	86.50	80.08	92.32	92.53
Robot	54.07	-	-	91.42	91.39
Foreground	52.02	-	-	92.19	92.11
Background	93.82	-	-	99.08	99.07

more versatile in practical applications than our previous work, where every moving pixel not identified as manipulator is automatically classified as grasped object.

2) *Dealing with Motion Stops:* To alleviate the constant dependency on motion for successful prediction we propose to add a *ConvLSTM* layer after the ASPP module and replace every *Conv2d* layer in the decoder with the respective recurrent counterpart (see also Fig. 2). Whenever both input images are alike (i.e., no observable motion between frames), the network still estimates manipulator and object given the memorized predictions of previous frames. We quantitatively analyze the decrease in mIoU after a motion stop in Fig. 5.

C. Ablation Study

We experiment with various placements of the correlation layer and a co-attention operation [26, 28] to explain our architectural design choices. In [28] co-attention is performed after the ASPP module, and we add an experiment with the co-attention block placed at the position of the best-performing correlation setting for fair comparison. In Tab. II we list the mIoU for the task of agnostic segmentation from motion (Sec. III-B) with our synthetic validation set (Sec. IV-A.1) and joint robot-object masks as test set (Sec. IV-C) after five epochs of training. We further show the number of parameters in million. For the inference estimate we average ten forward passes (single batch size, including image pre-processing) in evaluation mode. Our proposed setting of employing a correlation layer after the second ResNet block results in best performance on both data splits, while simultaneously being substantially faster than all other settings. We hypothesize that the comparatively bad performance of the co-attention layer stems from strong augmentations on both images.

D. Semantic Segmentation as Downstream Task

We evaluate the quality of our automatically generated object masks by employing them as training data for semantic segmentation. Specifically, we are interested in the performance of these masks in contrast to manually annotated

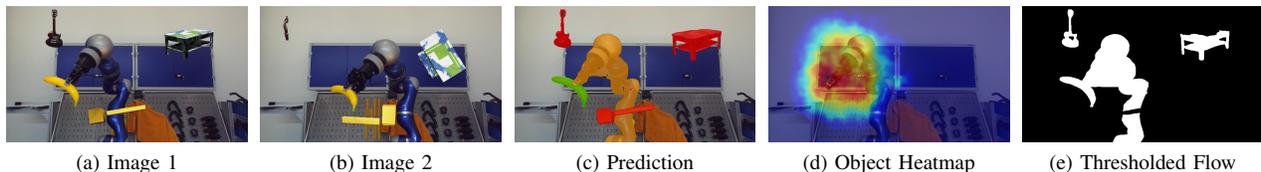


Fig. 4: Augmented moving background objects between two consecutive frames (a, b); Our network can directly separate moving foreground into grasped object (green), manipulator (orange) and other moving objects (red) (c); We sample 1,000 different objects on random positions on the image and visualize whether the item is identified as grasped object or belongs to moving background objects (d); While thresholded optical flow successfully segments moving foreground the binary prediction requires additional reasoning to differentiate objects into manipulator, object and distractor (e).

TABLE II: Architectural ablation study on different merge layers and positions. Our setting (**bold**) results in best performance for moving foreground segmentation in terms of mIoU on validation and test set while operating at ~ 14 Hz.

Merge Type	Pos. After	mIoU V/T [%]	MParams	Inf. [ms]
Correlation	ResBl-1	87.76 / 79.36	85.03	229.69
Correlation	ResBl-2	89.15 / 82.16	64.75	71.12
Correlation	ResBl-3	88.74 / 81.99	93.88	102.55
Correlation	ResBl-4	84.19 / 67.67	166.28	172.68
Correlation	ASPP	74.85 / 60.98	51.96	98.88
Co-Attention	ResBl-2	28.05 / 02.88	45.33	73.77
Co-Attention	ASPP	56.91 / 39.42	41.59	111.22
Co-Attention-Orig*	ASPP	50.61 / 20.02	41.59	115.75

*Here, segmentation maps for both inputs are predicted as per [28].

object masks (i.e., the ground truth annotations we compare to in Tab. I), and the improvements to our previous work.

Diverse training data (50,000 images) is generated by pasting object masks with random scale and translation on MS COCO [48] backgrounds. We train a standard DeepLabv3+ [39] with the same settings as for moving object segmentation (Sec. IV-B) and compare ground truth and predicted masks on images of STIOS [52] in Tab. III. We visualize exemplary results in Fig. 6.

For equal views across all masks the ground truth annotations represent an upper border, and we perform superior to masks generated by [10] due to better initial

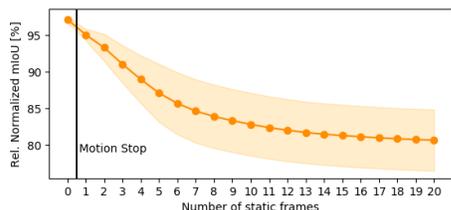


Fig. 5: After ten consecutive frames we simulate a motion stop by feeding equal input pairs to the network ($t=0$ denotes the last moving input pair). We repeat this five times per object, and plot mean and variance of the IoU relative to the point where motion last occurred across all objects. On average, DistinctNet still predicts 80% of the initial segmentation mask after 20 static frames.

TABLE III: Mean IoU on segmentation from tabletop scenes. Training data are a subset (45) or all (301) masks per grasped object. We use our generated annotations, predictions from [10] as well as ground truth annotations.

#Masks / Obj.	Ground Truth	[10]	Ours
45 Masks	58.91	50.39	56.79
301 Masks	-	54.78	59.86

quality (see Tab. I). Remarkably, the network benefits from many additional views, and employing all object views segmented by our approach as training data performs better than the smaller subset of ground truth annotations. Note that while there is no experiment with 301 ground truth masks per object we expect an analogous upper bound.

VI. CONCLUSION

We have presented a novel method for self-supervised grasped object segmentation from robotic manipulation. Although all our learning is based on observing motion in a static scene, a major contribution of this work is the identification and distinction of other moving background objects, as well as coping with stagnating motion. Our framework presents a reliable way of large-scale training data generation from real sensor environments, and the resulting masks are well suited for downstream tasks like semantic segmentation.

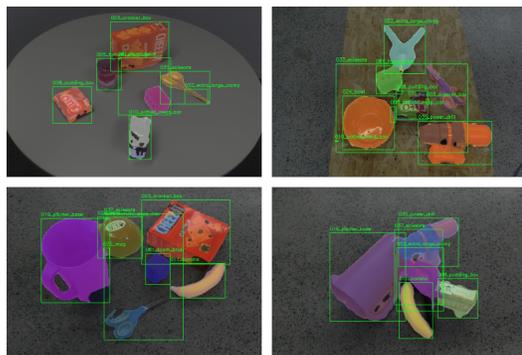


Fig. 6: Qualitative results from our semantic segmentation from scenes (best viewed magnified). We add bounding boxes to denote the class labels of the respective annotations.

REFERENCES

- [1] P. M. Fitzpatrick and G. Metta, "Grounding vision through experimental manipulation," *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 361, pp. 2165–2185, 2003. [Online]. Available: <https://royalsocietypublishing.org/doi/10.1098/rsta.2003.1251>
- [2] 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)*, May 2014, pp. 4959–4966. [Online]. Available: <https://ieeexplore.ieee.org/document/6907586>
- [3] D. Pathak, Y. Shentu, D. Chen, P. Agrawal, T. Darrell, S. Levine, and J. Malik, "Learning instance segmentation by interaction," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 2123–2133. [Online]. Available: <https://ieeexplore.ieee.org/document/8575444>
- [4] X. Deng, Y. Xiang, A. Mousavian, C. Eppner, T. Bretl, and D. Fox, "Self-supervised 6d object pose estimation for robot manipulation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 3665–3671.
- [5] M. Krainin, P. Henry, X. Ren, and D. Fox, "Manipulator and Object Tracking for In Hand Model Acquisition," *Proceedings, IEEE International Conference on Robots and Automation*, pp. 1817–1824, 2010. [Online]. Available: https://homes.cs.washington.edu/~xren/publication/krainin_icra10mm_in_hand_modeling.pdf
- [6] K. Welke, J. Issac, D. Schiebener, T. Asfour, and R. Dillmann, "Autonomous acquisition of visual multi-view object representations for object recognition on a humanoid robot," in *2010 IEEE International Conference on Robotics and Automation*. Anchorage, AK: IEEE, May 2010, pp. 2012–2019. [Online]. Available: <http://ieeexplore.ieee.org/document/5509328/>
- [7] B. Browatzki, V. Tikhanoff, G. Metta, H. H. Bulthoff, and C. Wallraven, "Active object recognition on a humanoid robot," in *2012 IEEE International Conference on Robotics and Automation*. St Paul, MN, USA: IEEE, May 2012, pp. 2021–2028. [Online]. Available: <http://ieeexplore.ieee.org/document/6225218/>
- [8] 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)*, 2019, pp. 8720–8726. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8794334>
- [9] V. Florence, J. J. Corso, and B. Griffin, "Robot-supervised learning for object segmentation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 1343–1349. [Online]. Available: <https://ieeexplore.ieee.org/document/9196543>
- [10] W. Boerdijk, M. Sundermeyer, M. Durner, and R. Triebel, "Self-Supervised Object-in-Gripper Segmentation from Robotic Motions," in *Conference on Robot Learning (CoRL)*, November 2020. [Online]. Available: https://corlconf.github.io/paper_275/
- [11] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 504–511, 2013. [Online]. Available: <https://ieeexplore.ieee.org/document/6256668>
- [12] H. Hirschmuller, "Stereo Processing by Semiglobal Matching and Mutual Information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328–341, Feb. 2008. [Online]. Available: <http://ieeexplore.ieee.org/document/4359315/>
- [13] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 1, pp. 185–203, 1981.
- [14] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *ECCV*, 2020. [Online]. Available: <https://arxiv.org/abs/2003.12039>
- [15] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 978–994, 2011. [Online]. Available: <https://ieeexplore.ieee.org/document/5551153>
- [16] I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional neural network architecture for geometric matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/papers/Rocco_Convolutional_Neural_Network_CVPR_2017_paper.pdf
- [17] J. Lee, D. Kim, J. Ponce, and B. Ham, "Sfnet: Learning object-aware semantic correspondence," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2273–2282. [Online]. Available: <https://ieeexplore.ieee.org/document/8954231>
- [18] R. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: a systematic survey," *IEEE Transactions on Image Processing*, vol. 14, no. 3, pp. 294–307, Mar. 2005. [Online]. Available: <https://ieeexplore.ieee.org/document/1395984>
- [19] T. Bouwmans, "Subspace Learning for Background Modeling: A Survey," *Recent Patent On Computer Science*, vol. 2, no. 3, pp. 223–234, Nov. 2009. [Online]. Available: <https://hal.archives-ouvertes.fr/hal-00534555>
- [20] 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*, vol. 117, pp. 8–66, Sept. 2019. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0893608019301303>
- [21] K. L. de Jong and A. Sergeevna Bosman, "Unsupervised change detection in satellite images using convolutional neural networks," in *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–8. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8851762>
- [22] L. Ru, B. Du, and C. Wu, "Multi-Temporal Scene Classification and Scene Change Detection with Correlation based Fusion," *arXiv:2006.02176 [cs]*, June 2020, arXiv: 2006.02176. [Online]. Available: <http://arxiv.org/abs/2006.02176>
- [23] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbaş, V. Golkov, P. v.d. Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2015. [Online]. Available: <http://lmb.informatik.uni-freiburg.de/Publications/2015/DFIB15>
- [24] T.-W. Hui, X. Tang, and C. C. Loy, "Liteflownet: A lightweight convolutional neural network for optical flow estimation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018, pp. 8981–8989. [Online]. Available: <https://arxiv.org/abs/1805.07036>
- [25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6000–6010. [Online]. Available: <https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>
- [26] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical Question-Image Co-Attention for Visual Question Answering," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 289–297. [Online]. Available: <https://arxiv.org/abs/1606.00061>
- [27] L. Wang, D. Liu, R. Puri, and D. N. Metaxas, "Learning trailer moments in full-length movies," 2020. [Online]. Available: <https://arxiv.org/abs/2008.08502>
- [28] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, "See More, Know More: Unsupervised Video Object Segmentation With Co-Attention Siamese Networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, June 2019, pp. 3618–3627. [Online]. Available: <https://ieeexplore.ieee.org/document/8953755/>
- [29] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV: IEEE, June 2016, pp. 724–732. [Online]. Available: <https://ieeexplore.ieee.org/document/7780454/>
- [30] A. Cioppa, M. V. Droogenbroeck, and M. Braham, "Real-time semantic background subtraction," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 3214–3218.
- [31] L. Sevilla-Lara, D. Sun, V. Jampani, and M. J. Black, "Optical Flow with Semantic Segmentation and Localized Layers," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, June 2016, pp. 3889–3898. [Online]. Available: <http://ieeexplore.ieee.org/document/7780791/>
- [32] M. Bai, W. Luo, K. Kundu, and R. Urtasun, "Exploiting semantic information and deep matching for optical flow," in *ECCV*, 2016. [Online]. Available: <https://arxiv.org/abs/1604.01827>

- [33] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Bideau_The_Best_of_CVPR_2018_paper.html
- [34] J. Hur and S. Roth, "Joint optical flow and temporally consistent semantic segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 163–177. [Online]. Available: <https://arxiv.org/abs/1607.07716>
- [35] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang, "SegFlow: Joint Learning for Video Object Segmentation and Optical Flow," in *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: IEEE, Oct. 2017, pp. 686–695. [Online]. Available: <http://ieeexplore.ieee.org/document/8237343/>
- [36] P. Tokmakov, C. Schmid, and K. Alahari, "Learning to segment moving objects," *International Journal of Computer Vision*, vol. 127, no. 3, pp. 282–301, 2019.
- [37] S. D. Jain, B. Xiong, and K. Grauman, "Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos," in *2017 IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, 2017, pp. 2117–2126.
- [38] "The KITTI Vision Benchmark Suite." [Online]. Available: http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark=flow
- [39] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 833–851. [Online]. Available: https://openaccess.thecvf.com/content_ECCV_2018/papers/Liang-Chieh-Chen.Encoder-Decoder.with.Atrous.ECCV_2018_paper.pdf
- [40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. [Online]. Available: <https://ieeexplore.ieee.org/document/7780459>
- [41] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.
- [42] G. Yang, H. Zhao, J. Shi, Z. Deng, and J. Jia, "SegStereo: Exploiting Semantic Information for Disparity Estimation," in *Computer Vision – ECCV, 2018*.
- [43] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, pp. 802–810. [Online]. Available: <https://arxiv.org/abs/1506.04214>
- [44] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "CDnet 2014: An Expanded Change Detection Benchmark Dataset," in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Columbus, OH, USA: IEEE, June 2014, pp. 393–400. [Online]. Available: <https://ieeexplore.ieee.org/document/6910011>
- [45] H. Abu Alhaija, S. K. Mustikovela, L. Mescheder, A. Geiger, and C. Rother, "Augmented Reality Meets Computer Vision: Efficient Data Generation for Urban Driving Scenes," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 961–972, Sept. 2018. [Online]. Available: <http://link.springer.com/10.1007/s11263-018-1070-x>
- [46] M. Denninger, M. Sundermeyer, D. Winkelbauer, Y. Zidan, D. Olefir, M. Elbadrawy, A. Lodhi, and H. Katam, "Blenderproc," *arXiv:1911.01911 [cs]*, Oct. 2019, arXiv: 1911.01911. [Online]. Available: <http://arxiv.org/abs/1911.01911>
- [47] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic Scene Completion from a Single Depth Image," *arXiv:1611.08974 [cs]*, Nov. 2016, arXiv: 1611.08974. [Online]. Available: <http://arxiv.org/abs/1611.08974>
- [48] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755. [Online]. Available: <https://arxiv.org/abs/1405.0312>
- [49] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "Shapenet: An information-rich 3d model repository," Stanford University — Princeton University — Toyota Technological Institute at Chicago, Tech. Rep. arXiv:1512.03012 [cs.GR], 2015. [Online]. Available: <https://arxiv.org/abs/1512.03012>
- [50] D. Dwibedi, I. Misra, and M. Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1310–1319. [Online]. Available: <https://ieeexplore.ieee.org/document/8237408>
- [51] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *arXiv:1711.05101 [cs, math]*, Jan. 2019, arXiv: 1711.05101. [Online]. Available: <http://arxiv.org/abs/1711.05101>
- [52] M. Durner, W. Boerdijk, M. Sundermeyer, W. Friedl, Z.-C. Marton, and R. Triebel, "Unknown object segmentation from stereo images," 2021, arXiv: 2103.06796. [Online]. Available: <https://arxiv.org/abs/2103.06796>