Video Object Segmentation using Teacher-Student Adaptation in a Human Robot Interaction (HRI) Setting

Mennatullah Siam[1], Chen Jiang[1], Steven Lu[1], Laura Petrich[1],
Mahmoud Gamal[2], Mohamed Elhoseiny[3], Martin Jagersand[1]

Abstract— Video object segmentation is an essential task in robot manipulation to facilitate grasping and learning affordances. Incremental learning is important for robotics in unstructured environments. Inspired by the children learning process, human robot interaction (HRI) can be utilized to teach robots about the world guided by humans similar to how children learn from a parent or a teacher. A human teacher can show potential objects of interest to the robot, which is able to self adapt to the teaching signal without providing manual segmentation labels. We propose a novel teacher-student learning paradigm to teach robots about their surrounding environment. A two-stream motion and appearance "teacher" network provides pseudo-labels to adapt an appearance "student" network. The student network is able to segment the newly learned objects in other scenes, whether they are static or in motion. We also introduce a carefully designed dataset that serves the proposed HRI setup, denoted as (I)nteractive (V)ideo (O)bject (S)egmentation. Our IVOS dataset contains teaching videos of different objects, and manipulation tasks. Our proposed adaptation method outperforms the state-of-the-art on DAVIS and FBMS with 6.8% and 1.2% in F-measure respectively. It improves over the baseline on IVOS dataset with 46.1% and 25.9% in mIoU.

I. INTRODUCTION

The robotics and vision communities greatly improved video object segmentation over the recent years. The main approaches in video object segmentation could be categorized into semi-supervised and unsupervised approaches. In semi-supervised video object segmentation approaches (e.g., [32][2][12], the method is initialized manually by a segmentation mask in the first few frames, then the segmented object is tracked throughout the video sequence. On the other hand, unsupervised methods [14][30][9][29] attempt to discover the primary object automatically and segment it through the video sequence. Motion is one of the fundamental cues that can help improve unsupervised video object segmentation. While there has been recent success in deep learning approaches for segmenting motion (e.g., [30][9][29]), current approaches depend mainly on prior large-scale training data.

Video semantic segmentation for robotics is widely used in different applications such as autonomous driving [4][24], and robot manipulation [5][10]. Object segmentation can aid in grasping, manipulating objects, and learning object affordances [5]. In robot manipulation, learning to segment new
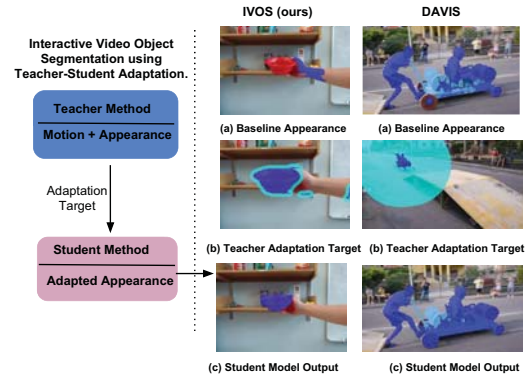

Fig. 1: Overview of the proposed Teacher-Student adaptation method for video object segmentation. The teacher model based on motion cues is able to provide pseudo-labels to adapt the student model. Blue: confident positive pixels. Cyan: ignored region in the adaptation.

objects incrementally, has significant importance. Real world environments have far more objects and more appearance variation than can be feasibly trained a-priori. Current large-scale datasets such as Image-Net [15] do not cover this.

A recent trend in robotics is toward human-centered artificial intelligence. Human-centered AI involves learning by instruction using a human teacher. Such human-robot interaction (HRI) mimics children being taught novel concepts from few examples [18]. In the robotic setting, a human teacher demonstrates an object by moving it and showing different poses, while verbally or textually teaching its label. The robot is then required to segment the objects in other settings where it is either static or manipulated by the human or the robot itself. We demonstrated this HRI setting in our team submission to the KUKA Innovation Challenge at the Hannover Fair [23]. This HRI setting has few differences to conventional video object segmentation: (1) Abundance of the different poses of the object. (2) The existence of different instances/classes within the same category. (3) Different challenges introduced by cluttered backgrounds, different rigid and non-rigid transformations, occlusions and illumination changes. In this paper, we focus on these robotics challenges and provide a new dataset and a new method to study such a scenario.

We collected a new dataset to benchmark (I)nteractive (V)ideo (O)bject (S)egmentation in the HRI scenario. The dataset contains two types of videos: (1) A human teacher

[1]Mennatullah Siam, Chen Jian, Steven Lu, Laura Petrich and Martin Jagersand are with the University of Alberta, Canada. e-mail: mennatul@ualberta.ca.

[2]Mahmoud Gamal is with Cairo University, Egypt.

[3] Mohamed El-Hoseiny is with Facebook AI Research. e-mail: elhoseiny@fb.com.

showing different household objects in varying poses for interactive learning. (2) Videos of the same objects used in a kitchen setting while serving and eating food. The objects occur both as static objects and active objects being manipulated. Manipulation was performed by both humans and robots. The aim of this dataset is to facilitate incremental learning and immediate use in a collaborative human-robot environments, such as assistive robot manipulation. Datasets that have a similar setting such as ICUBWorld transformations dataset [21], and the Core50 dataset [17] were proposed. These datasets include different instances within the same category. They benchmark solutions to object recognition in a similar HRI setting but do not provide segmentation annotations unlike our dataset. Other datasets were concerned with the activities of daily living such as the ADL dataset [22]. The dataset was comprised of ego-centric videos for activities. However, such ADL datasets do not contain the required teaching videos to match the HRI setting we are focusing on. Table I summarizes the most relevant datasets suited to the HRI setting.

The main contribution of our collected IVOS dataset is providing the manipulation tasks setting with objects being manipulated by humans or a robot. In addition to providing segmentation annotation for both teaching videos and manipulation tasks. It enables researchers to analyze the effect of different transformations such as translation, scale, and rotation on the incremental learning of video object segmentation. It acts as a benchmark for interactive video object segmentation in the HRI setting. It also provides videos of both human and similarly robot manipulation tasks with the segmentation annotations along with the corresponding robot trajectories. Thus, it enables further research in learning robot trajectories from visual cues with semantics.

We propose a novel teacher-student adaptation method based on motion cues for video object segmentation. Our method enables a human teacher to demonstrate objects moving with different transformations and associates them with labels. During inference, our approach can learn to segment the object without manual segmentation annotation. The teacher model is a fully convolutional network that combines motion and appearance, denoted as "Motion+Appearance". The adapted student model is a one-stream appearance-only fully convolutional network denoted as "Appearance". Combining motion and appearance in the teacher network allows the creation of pseudo-labels for adapting the student network. Our work is inspired from the semi-supervised on-line method [32]. This work uses manual segmentation masks for initialization. Instead, our approach tackles a more challenging problem and does not require manual segmentation; it relies on the pseudo-labels provided by the teacher model. Figure 1 shows an overview of the proposed method. The two main reasons behind using the adaptation targets from the teacher model is: (1) The student model is more computationally efficient. The inference and adaptation time for the teacher model is 1.5x of the student model's. The adaptation occurs only once on the first frame, then the more efficient student model can be used for inference. (2)

TABLE I: Comparison of different datasets. T:Turntable, H:handheld

Dataset	Sess.	Cat.	Obj.	Acq.	Tasks	Seg.
RGB-D [25]	-	51	300	T	✗	✗
BIG BIRD [26]	-	-	100	T	✗	✗
ICUB 28 [20]	4	7	28	H	✗	✗
ICUB World [21]	6	20	200	H	✗	✗
Core50 [17]	11	10	50	H	✗	✗
IVOS	12	12	36	H	✓	✓

The teacher model can be used to generate pseudo-labels for the potential object of interest. It does not require the human to provide manual segmentation mask during the teaching phase which provides a natural interface to the human. Consequently, the adapted student model can segment the object of the interest whether it is static or moving. If the adapted model was still dependant on optical flow it will only be able to recognize the object in motion.

Our proposed method outperforms the state-of-the-art on the popular DAVIS and FBMS benchmarks with 6.8% and 1.2% in F-measure respectively. On our new IVOS dataset results show the motion adapted network outperforms the baseline with 46.1% and 25.9% in mIoU on Scale/Rotation and Manipulation Tasks respectively. Our code [1] and IVOS dataset [2] are publicly available. A video description and demonstration is available at [3]. Our main contributions are :

- Providing a Dataset for Interactive Video Object Segmentation (IVOS) in a Human-Robot Interaction setting, and including manipulation tasks unlike previous datasets.
- A teacher-student adaptation method is proposed to learn new objects from a human teacher without providing manual segmentation labels. We propose a novel pseudo-label adaptation based on a teacher model that is dependant on motion. Adaptation with discrete and continuous pseudo-labels are evaluated to demonstrate different adaptation methods.

II. IVOS DATASET

We collected IVOS for the purpose of benchmarking (I)nteractive (V)ideo (O)bject (S)egmentation in the HRI setting. We collect the dataset in two different settings: (1) Human teaching objects. (2) Manipulation tasks setting. Unlike previous datasets in human robot interaction IVOS dataset provides video sequences for manipulation tasks. In addition to providing segmentation annotation for both teaching videos and manipulation tasks.

A. Human Teaching Objects

For teaching, videos are collected while a human moves an object with her hand. The unstructured human hand motion naturally provides different views of the object and samples different geometric transformations. We provide transformations such as translation, scale, planar rotation, out-of-plane

[1] https://github.com/MSiam/motion_adaptation
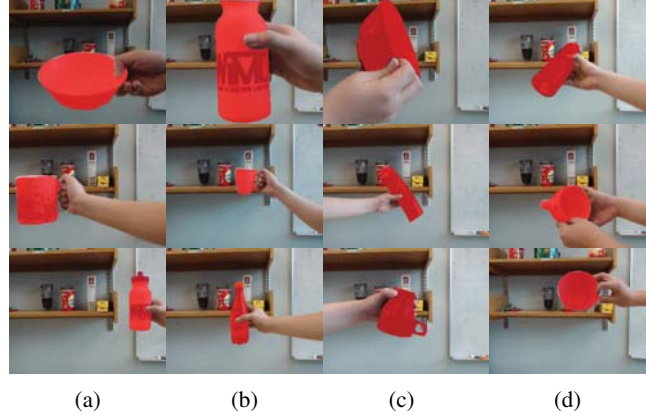[2] https://msiam.github.io/ivos/
[3] https://youtu.be/36hMbAs8e0c

Fig. 2: Samples of collected Dataset IVOS, Teaching Objects Setting. (a) Translation split. (b) Scale split. (c) Planar Rotation split. (d) Out-of-plane Rotation. (e) Non rigid transformations.

Fig. 3: Samples of collected IVOS dataset, Robot manipulation Tasks Setting with segmentation annotation. Manipulation Tasks: (a) Drinking. (b) Stirring. (c) Pouring Water. (d) Pouring Cereal.

rotation, and other transformations such as opening the lid of a bottle. Two illumination conditions are provided: daylight and indoor lighting, which sums up to 10 sessions of recording for both illumination and transformations. Figure 2 shows a sample for the objects being captured under different transformations with the segmenation masks. In each session a video for the object held by a human with relatively cluttered scene background is recorded.

A GRAS-20S4C-C fire-wire camera is used to record the data along with a Kinect sensor [27]. The collected data is annotated manually with polygonal masks using the VGG Image Annotator tool [6]. The final teaching videos contains 12 object categories, with a total of 36 instances under these categories. The detection crops are provided for all the frames, while the segmentation masks are provided for 20 instances with $\sim$ 18,000 annotated masks.

B. Manipulation Tasks Setting

The manipulation task benchmark includes two video categories: one with human manipulation, and the other with robot manipulation. Activities of Daily Living (ADL) such as food preparation are the focus for the recorded tasks. The aim of this benchmark is to further improve perception systems in robotics for assisted living. Robot trajectories are created through kinesthetic teaching, and the robot pose way-points are provided in the dataset. In order to create typical robot velocity and acceleration, profiles trajectories

were generated from these way-points using splines as is standard in robotics.

The collected sequences are further annotated with segmentation masks similar to the teaching objects setting. Figure 3 shows some of the recorded frames with ground-truth annotations. It covers 4 main manipulation tasks: *cutting, pouring, stirring, and drinking* for both robot and human manipulation covering a total of 56 tasks. The dataset contains $\sim$ 8,900 frames with segmentation masks, along with the recorded robot trajectories to enable further research on how to learn these trajectories from visual cues.

III. METHOD

A. Baseline Network Architecture

The student model in this work is built on the wide ResNet architecture presented in [34]. The network is comprised of 16 residual blocks. Dilated convolution [35] is used to increase the receptive field without decreasing the resolution. The output from the network is bilinearly upsampled to the initial image resolution. The loss function used is bootstrapped cross entropy [33], which helps with class imbalance. It computes the cross entropy loss from a fraction of the hardest pixels. Pre-trained weights on PASCAL dataset for objectness is used from [32], to help the network generalize to different objects in the scene. Then it is trained on DAVIS training set, the student model without adaptation is denoted as the baseline model throughout the paper.
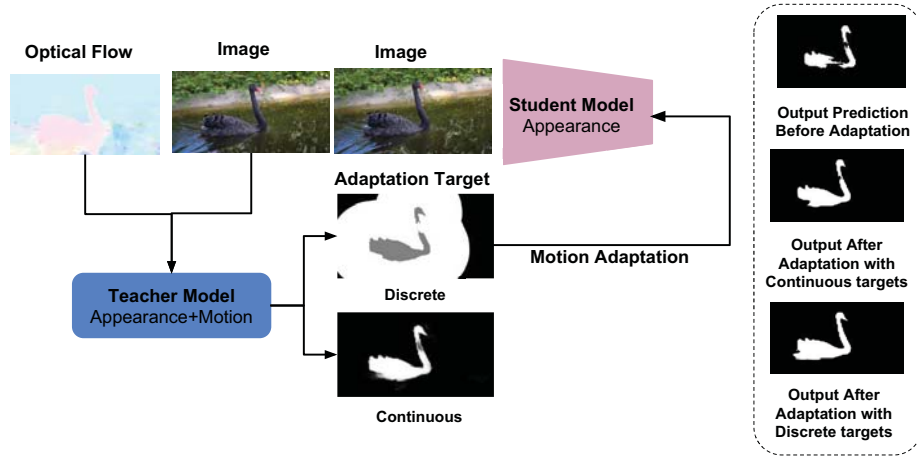
Fig. 4: Motion Adaptation of fully convolutional residual networks pipeline.

The teacher network incorporates motion from optical flow, where a two-stream wide ResNet for motion and appearance is used. Each stream contains 11 residual blocks for memory efficiency reasons. The output feature maps are combined by multiplying the output activation maps from both motion and appearance streams. After combining features another 5 residual blocks are used with dilated convolution. The input to the motion stream is the optical flow computed using [16], and converted into RGB representation using the Sintel color wheel representation [1].

B. Teacher-Student Adaptation using Pseudo-labels

There is an analogy between this work and the work in [31], where a student method is learning to mimic a teacher method. In our work the teacher method is a motion dependent one, and the student method tries to mimic the teacher during inference through motion adaptation. The teacher-student training helps the network understands the primary object in the scene in an unsupervised manner. Unlike the work in [32] that first fine-tunes the network based on the manual segmentation mask then adapts it online with the most confident pixels. Our method provides a natural human robot interaction that does not require manual labelling for initialization.

Our approach provides two different adaptation methods, adapting based on discrete or continuous labels. The teacher network pseudo-labels are initially filtered to remove parts representing the human moving using the output human segmentation from Mask R-CNN [8]. When discrete labels are used it is based on pseudo-labels from the confident pixels in the teacher network output. Such a method provides superior accuracy, but on the expense of tuning the parameters that determine these confident pixels. Another method that utilizes continuous labels adaptation from the teacher network is also introduced. This method alleviates the need for any hyper-parameter tuning but on the cost of degraded accuracy. Figure 4 summarizes the adaptation scheme, and shows the output pseudo-labels, the output segmentation before and after adaptation.

In the case of discrete pseudo-labels, the output probability maps from the teacher network is further processed in a similar fashion to the semi-supervised method [32]. Initially the confident positive pixels are labeled, then a geometric distance transform is computed to label the most confident negative pixels as shown in Algorithm 1.

Algorithm 1 Motion Adaptation Algorithm.
Input: X: images used for teaching. N: number of samples used. $M_{teacher}$: Teacher Model. $M_{student}$: Student Model.
Output: $\hat{M}_{student}$: Adapted Student Model.

1: **function** TEACH(N, X, $M_{teacher}$, $M_{student}$)
2:     **for** i in N **do**
3:         $P_i = M_{teacher}(X_i)$
4:         $\hat{M}_{student}$ = Adapt(P_i, $M_{student}$)
5:     **end for**
6: **end function**
    Discrete Labels Adaptation Method
7: **function** ADAPT(A_t, $M_{student}$)
8:     Mask ← IGNORED
9:     pos_indices ← (A_t > POS_TH)
10:     dt ← DITANCE_TRANSFORM(Mask)
11:     neg_indices ← (dt > NEG_DT_TH)
12:     Mask[pos_indices] ← 1, Mask[neg_indices] ← 0
    return finetune($M_{student}$,Mask)
13: **end function**

In the case of continuous labels, the output probability maps are used without further processing. This has the advantage of not using any hyper-parameters or discrete label segmentation. It generalizes better to different scenarios on the expense of degraded accuracy. Inspiring from the relation between cross entropy and KL-divergence as in equations 1. The cross entropy loss can be viewed as a mean to decrease the divergence between the true distribution p and the predicted one q, in addition to the uncertainty implicit in $H(p)$. In our case the true distribution is the probability maps from the teacher network, while the predicted is the student

Authorized licensed use limited to: UNIVERSITY OF ALBERTA. Downloaded on June 13,2023 at 20:55:06 UTC from IEEE Xplore. Restrictions apply.

network output. Figure 5 shows the difference between the pseudo-labels for both discrete and continuous variants. Conditional random fields is used as a post-processing step on DAVIS and FBMS.
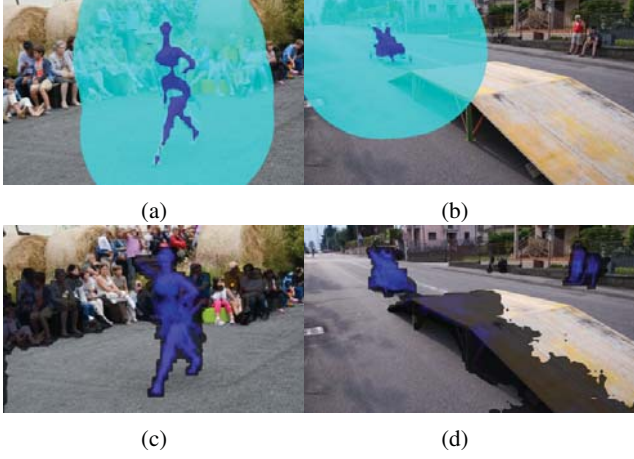

| (a) | (b) |
| (c) | (d) |

Fig. 5: (a,b) Discrete adaptation targets (pseudo-labels), cyan is the unknown region, blue is the confident positive pixels. (c, d) Continuous adaptation targets.

$$D_{KL}(p|q) = \sum_i p_i \log \frac{p_i}{q_i} \qquad (1a)$$

$$D_{KL}(p|q) = \sum_i p_i \log \frac{1}{q_i} - H(p) \qquad (1b)$$

$$H(p,q) = H(p) + D_{KL}(p|q) \qquad (1c)$$

IV. EXPERIMENTAL RESULTS

A. Experimental Setup

For all experiments the DAVIS training data is used to train our Appearance model and the Appearance+Motion model. The optimization method used is Adam [13] with learning rate 10^{-6} during training, and 10^{-5} during on-line adaptation. In on-line adaptation 15 iterations are used in the scale/rotation experiments and 50 in the tasks experiments. Adaptation is only conducted once at the initialization of the video object segmentation. The positive threshold used to identify highly confident positive samples is 0.8, and the negative threshold distance to the foreground mask is 220 in case of DAVIS benchmark. Since IVOS is recorded in an indoor setup, a negative distance threshold of 20 is used.

B. Generic Video Object Segmentation

In order to evaluate the performance of our proposed motion adaptation (MotAdapt) method with respect to the state-of-the-art, we experiment on generic video object segmentation datasets. Table II shows quantitative analysis on DAVIS benchmark compared to the state-of-the-art unsupervised methods. One of the variants of MotAdapt based on discrete labels outperforms the state of the art with 6.8% in

F-measure, and 1% in mIoU. Table III shows quantitative results on FBMS dataset, where our MotAdapt outperforms the state of the art with 1.2% in F-measure and 10% in recall.

Figure 6 shows qualitative results on FBMS highlighting the improvement gain from motion adaptation compared to LVO [30]. Figure 7 shows qualitative evaluation on DAVIS, where it demonstrates the benefit from motion adaptation compared to the baseline (top row), and compared to LVO [30] and ARP [14] (bottom row).

C. Video Object Segmentation in HRI Setting

Our method is evaluated in the HRI scenario on our dataset IVOS. The teaching is performed on the translation sequences, with only the first two frames used to generate pseudo-labels for adaptation. An initial evaluation is conducted on both scale and rotation sequences, in order to assess the adaptation capability to generalize to different poses and transformations. Table IV shows the comparison between the baseline method without adaptation, and the two variants of motion adaptation on the scale, rotation and tasks sequences. The discrete and continuous variants for our motion adaptation outperform the baseline with 54.5% and 49.3% respectively on the scale sequences. Similarly on the rotation sequences it outperforms the baseline with 37.7% and 35.7% respectively. The main reason for this large gap, is that general segmentation methods will segment all objects in the scene as foreground, while our teaching method adaptively learns the object of interest that was demonstrated by the human teacher.

All manipulation tasks sequences where the category bottle existed is evaluated and cropped to include the working area. Our method outperforms the baseline on the tasks with 25.9%. The first variant of our adaptation method generally outperforms the second variant with continuous labels adaptation. However the second variant has the advantage that it can work on any setting such as DAVIS and IVOS without tuning any hyper-parameters. Figure 8 shows the output from our adaptation method when it is recognized by the robot, and while the robot has successfully manipulated that object.

V. CONCLUSIONS

In this paper we proposed a novel approach for visual learning by instruction. Our proposed motion adaptation (MotAdapt) method provides a natural interface to teaching robots to segment novel object instances. This enables robots to manipulate and grasp these objects. Two variants of the adaptation scheme is experimented with. Our results show that Mot-Adapt outperforms the state of the art on DAVIS and FBMS and outperforms the baseline on IVOS dataset.

REFERENCES

[1] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *International Journal of Computer Vision*, vol. 92, no. 1, pp. 1–31, 2011.
[2] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," *arXiv preprint arXiv:1611.05198*, 2016.
[3] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang, "Segflow: Joint learning for video object segmentation and optical flow," *arXiv preprint arXiv:1709.06750*, 2017.

TABLE II: Quantitative comparison on DAVIS benchmark. MotAdapt-1: Continuous Labels, MotAdapt-2: Discrete Labels.

Measure		NLC[7]	SFL[3]	LMP [29]	FSeg [9]	LVO [30]	ARP [14]	Baseline	MOTAdapt-1	MOTAdapt-2
$\mathcal{J}$	Mean	55.1	67.4	70.0	70.7	75.9	76.2	74.0	75.3	**77.2**
	Recall	55.8	81.4	85.0	83.5	89.1	**91.1**	85.7	87.1	87.8
	Decay	12.6	6.2	1.3	1.5	**0.0**	**0.0**	7.0	5.0	5.0
$\mathcal{F}$	Mean	52.3	66.7	65.9	65.3	72.1	70.6	74.4	75.3	**77.4**
	Recall	51.9	77.1	79.2	73.8	83.4	83.5	81.6	83.8	**84.4**
	Decay	11.4	5.1	2.5	1.8	1.3	7.9	**0.0**	3.3	3.3

TABLE III: Quantitative results on FBMS dataset (test set).

Measure	FST [19]	CVOS [28]	CUT [11]	MPNet-V[29]	LVO[30]	Base	ours
$\mathcal{P}$	76.3	83.4	83.1	81.4	**92.1**	80.8	80.7
$\mathcal{R}$	63.3	67.9	71.5	73.9	67.4	76.1	**77.4**
$\mathcal{F}$	69.2	74.9	76.8	77.5	77.8	78.4	**79.0**

TABLE IV: mIoU on IVOS over the different transformations and tasks. IVOS dataset teaching is conducted on few samples from the translation, then evaluating on scale, rotation and manipulation tasks. MotAdapt-1: Continuous Labels. MotAdapt-2: Discrete Labels.

Model	Scale	Rotation	Manipulation Tasks
Baseline	14.5	13.8	14.7
MotAdapt-1	63.8	49.5	30.2
Mot-Adapt-2	**69.0**	**51.5**	**40.6**

Fig. 6: Qualitative Evaluation on the FBMS dataset. Top: LVO [30]. Bottom: ours.
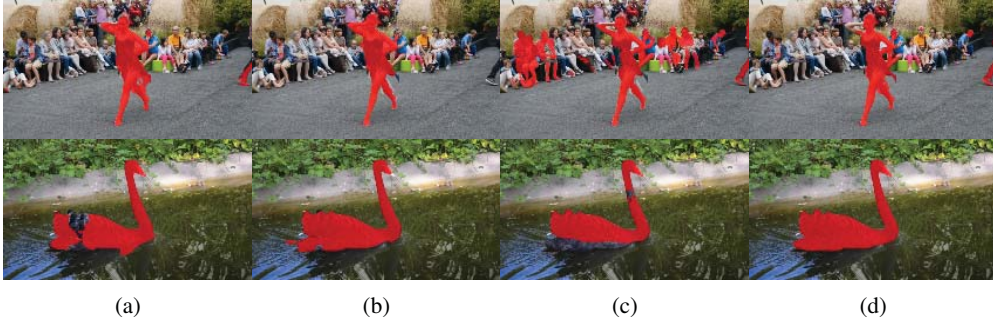

(a) (b) (c) (d)

Fig. 7: Qualitative evaluation on DAVIS16. (a) LVO [30]. (b) ARP [14]. (c) Baseline. (d) MotAdapt.

(a) (b) (c) (d)

Fig. 8: Qualitative evaluation on IVOS Manipulation Tasks Setting. (a) Teaching Phase, Discrete Labels. (b) Teaching Phase, Continuous Labels. (c) Inference Phase before manipulation. (d) Inference Phase, during manipulation.

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[5] T.-T. Do, A. Nguyen, I. Reid, D. G. Caldwell, and N. G. Tsagarakis, "Affordancenet: An end-to-end deep learning approach for object affordance detection," *arXiv preprint arXiv:1709.07326*, 2017.

[6] A. Dutta, A. Gupta, and A. Zissermann, "VGG image annotator (VIA)," http://www.robots.ox.ac.uk/ vgg/software/via/, 2016.

[7] A. Faktor and M. Irani, "Video segmentation by non-local consensus voting." in *BMVC*, vol. 2, no. 7, 2014, p. 8.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.

[9] S. D. Jain, B. Xiong, and K. Grauman, "Fusionseg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos," *arXiv preprint arXiv:1701.05384*, 2017.

[10] J. Kenney, T. Buckley, and O. Brock, "Interactive segmentation for manipulation in unstructured environments," in *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*. IEEE, 2009, pp. 1377–1382.

[11] M. Keuper, B. Andres, and T. Brox, "Motion trajectory segmentation via minimum cost multicuts," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3271–3279.

[12] A. Khoreva, F. Perazzi, R. Benenson, B. Schiele, and A. Sorkine-Hornung, "Learning video object segmentation from static images," *arXiv preprint arXiv:1612.02646*, 2016.

[13] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[14] Y. J. Koh and C.-S. Kim, "Primary object segmentation in videos based on region augmentation and reduction."

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[16] C. Liu *et al.*, "Beyond pixels: exploring new representations and applications for motion analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[17] V. Lomonaco and D. Maltoni, "Core50: a new dataset and benchmark for continuous object recognition," *arXiv preprint arXiv:1705.03550*, 2017.

[18] E. M. Markman, *Categorization and naming in children: Problems of induction*. Mit Press, 1989.

[19] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1777–1784.

[20] G. Pasquale, C. Ciliberto, F. Odone, L. Rosasco, and L. Natale, "Teaching icub to recognize objects using deep convolutional neural networks," in *Machine Learning for Interactive Systems*, 2015, pp. 21–25.

[21] G. Pasquale, C. Ciliberto, L. Rosasco, and L. Natale, "Object identification from few examples by improving the invariance of a deep convolutional neural network," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. IEEE, 2016, pp. 4904–4911.

[22] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2847–2854.

[23] K. Robotics, "KUKA Innovation Award Challenge," https://www.youtube.com/watch?v=aLcw73dt_Oo, 2018.

[24] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3234–3243.

[25] M. Schwarz, H. Schulz, and S. Behnke, "Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1329–1335.

[26] A. Singh, J. Sha, K. S. Narayan, T. Achim, and P. Abbeel, "Bigbird: A large-scale 3d database of object instances," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014, pp. 509–516.

[27] J. Steward, D. Lichti, J. Chow, R. Ferber, and S. Osis, "Performance assessment and calibration of the kinect 2.0 time-of-flight range camera for use in motion capture applications," in *Proceedings of the Fig Working Week*, 2015.

[28] B. Taylor, V. Karasev, and S. Soatto, "Causal video object segmentation from persistence of occlusions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4268–4276.

[29] P. Tokmakov, K. Alahari, and C. Schmid, "Learning motion patterns in videos," *arXiv preprint arXiv:1612.07217*, 2016.

[30] ——, "Learning video object segmentation with visual memory," *arXiv preprint arXiv:1704.05737*, 2017.

[31] G. Urban, K. J. Geras, S. E. Kahou, O. Aslan, S. Wang, R. Caruana, A. Mohamed, M. Philipose, and M. Richardson, "Do deep convolutional nets really need to be deep and convolutional?" *arXiv preprint arXiv:1603.05691*, 2016.

[32] P. Voigtlaender and B. Leibe, "Online adaptation of convolutional neural networks for video object segmentation," *arXiv preprint arXiv:1706.09364*, 2017.

[33] Z. Wu, C. Shen, and A. v. d. Hengel, "Bridging category-level and instance-level semantic image segmentation," *arXiv preprint arXiv:1605.06885*, 2016.

[34] ——, "Wider or deeper: Revisiting the resnet model for visual recognition," *arXiv preprint arXiv:1611.10080*, 2016.

[35] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.