A Weakly Supervised Strategy for Learning Object Detection on a Humanoid Robot

Elisa Maiettini[1,2,3] Giulia Pasquale[1,2] Vadim Tikhanoff[4] Lorenzo Rosasco[2,3] Lorenzo Natale[1]

***Abstract*— Research in Computer Vision and Deep Learning has recently proposed numerous effective techniques for detecting objects in an image. In general, these employ deep Convolutional Neural Networks trained end-to-end on large datasets annotated with object labels and 2D bounding boxes. These methods provide remarkable performance, but are particularly expensive in terms of training data and supervision. Hence, modern object detection algorithms are difficult to be deployed in robotic applications that require on-line learning. In this paper, we propose a weakly supervised strategy for training an object detector in this scenario. The main idea is to let the robot iteratively grow a training set by combining autonomously annotated examples, with others that are requested for human supervision. We evaluate our method on two experiments with data acquired from the iCub and R1 humanoid platforms, showing that it significantly reduces the number of human annotations required, without compromising performance. We also show the effectiveness of this approach when adapting the detector to a new setting.**

I. INTRODUCTION

State-of-the-art methods for object detection (the task of recognizing and localizing with a 2D bounding box every known object in an image) offer a variety of well-established deep learning tools to achieve high performance in challenging real world scenarios. These approaches generally rely on architectures trained end-to-end on datasets carefully collected and annotated (once and off-line). While this provides an effective baseline, considering the deployment on a humanoid robot to unconstrained environments, the adaptation capability is equally important. This includes learning to recognize novel, specific object instances, as well as tuning to specific settings, by relying on data gathered during the robot's operation ("on-line"), which may be scarce or not annotated. Moreover, the training may be constrained in terms of computational resources and time. In this paper we focus therefore on the problem of training and in particular *adapting* object detectors on-line on little, partially annotated data. We build on our previous work [1], [2], where we proposed a method to train a humanoid robot to detect novel object instances with training time in the order of seconds and only a few hundred frames. In [1], [2], however, supervision originated from interaction with a human teacher, while generalization to different background and light conditions was limited by the small number of training examples.

In this work we propose a strategy that allows the robot to adapt an object detector by acquiring new training samples with limited human intervention. The main idea is that, when faced with a new setting, the robot can iteratively adapt the object detector by parsing incoming images and either annotating them autonomously or asking for human help. This weakly supervised strategy integrates the fast object detector proposed in [2] with an adapted version of the Self-Supervised Sample Mining, SSM [3], [4].

As a benchmark, we rely on the publicly available iCub-World Transformations (iCWT) dataset [5], [6], that represents 200 objects handheld by a human interacting with the iCub [7]. To evaluate our method, we contribute an extension to the dataset, that represents 21 of its objects, randomly positioned in two table-top conditions, acquired with the R1 humanoid robot [8]. While our contribution provides a method to adapt detectors in scenarios where automatic annotation is challenging, for the sake of performance assessment, we chose a table top setting, which is distinct from the training procedure (refer to Fig. 2) but allowed us to automatically collect the ground-truth.

The resulting method shows successful results and allows our detection models to adapt to the new conditions, while limiting the amount of novel annotated images. The rest of the paper is organized as follows: section II overviews related work; section III describes our pipeline in detail and section IV presents results from the considered benchmark; finally, section V draws conclusions and outlines important directions for future work.

II. RELATED WORK

In this section we first overview recent methods for object detection in robotics (Sec. II-A), then we consider related works exploring the field of weakly supervised learning for object detection (Sec. II-B)

A. Object Detection for Robotic Applications

A major objective of latest research in object detection for robotics is to improve performance in difficult scenarios, targeting, e.g., occlusions and clutter [9], [10], [11], [12]. This is also reflected in challenges like the APC (Amazon Picking Challenge) [1]. To this end, a major trend is to rely on deep learning architectures, that can be stunningly effective in complex settings. Deep learning approaches can be grouped

[1] Humanoid Sensing and Perception, Istituto Italiano di Tecnologia, Genoa, Italy

[2] Laboratory for Computational and Statistical Learning, Istituto Italiano di Tecnologia and Massachusetts Institute of Technology, Cambridge, MA

[3] Dipartimento di Informatica, Bioingegneria, Robotica e Ingegneria dei Sistemi, University of Genoa, Genoa, Italy

[4] iCub Facility, Istituto Italiano di Tecnologia, Genoa, Italy

[1] http://amazonpickingchallenge.org/

into grid-based and region-based methods. Architectures in the first group typically apply a set of classifiers over a fixed, dense grid of locations in an image (see, e.g., SSD (Single-Shot MultiBox Detector) [13] and YOLO (You Only Look Once) [14], [15]). Methods in the second group, instead, consider for classification, a previously selected set of *region proposals* –i.e., regions which might contain objects of interest (see, e.g., Region-CNN (R-CNN) [16] and its evolutions Fast R-CNN [17], Faster R-CNN [18], Region-FCN [19] and Mask R-CNN [20]). In both groups, performance is usually achieved mostly through the collection of huge datasets, which require, long, time consuming training. In fact, the common trend is to combine the different stages of the typical object detection pipeline into a single model, that can be learned end-to-end via backpropagation [15], [19], [20], [18].

Nevertheless, we argue that a multi-stage architecture, learning each stage separately (see, e.g., [16], [17] and, specifically, [2]), might allow for faster strategies of adaptation, which is a critical requirement for many robotic applications.

Moreover, since the amount of possible locations in an image that might contain an object of interest (and that consequently need to be visited and classified) is typically large, the task of object detection is computationally heavy per-se. Considering that the majority of these regions then typically depicts background areas, the associated classification problem must be treated properly in order to avoid learning a biased predictor. To this end, solutions proposed in the literature are based either on (i) specific loss functions, to down-weight the contribution of the easier negative examples in the total loss (see, e.g., [21]), or on (ii) the idea of training a detector on a bootstrapped subset of harder background examples (see, e.g., [22], [23], [16], [24]).

In this work, we build on the multi-stage architecture proposed in [2]. This is composed of a deep learning-based region proposal and feature extractor (namely, a part of Faster R-CNN [18]), followed by a Kernel method for classification and a boostrapping approach to address the background-foreground imbalance. This pipeline is suited for a typical unconstrained robotic setting, as the combination of an extremely efficient classifier (FALKON [25]) with an approximated bootstrapping (see [2]), provides fast model training.

B. Weakly Supervised Learning for Object Detection

Gathering the ground truth for training object detection algorithms through supervised learning is a costly operation, since it requires drawing a bounding box around each object of interest (and provide its label) in each image example – and typically thousands are required.

While one approach that is gaining momentum is to rely on synthetic imagery [10], [26], the scope of this work is to consider latest research that focused on reducing this effort by adopting weakly or self-supervised (SS) techniques to extract as much information as possible from unlabeled or partially labeled images.

Methods to leverage on datasets annotated only at the image level (i.e., without bounding box information) were proposed to, respectively, learn an object detection system [27] or a region proposal generation algorithm [28].

Differently from applications where the images come from the web and no prior information about them is known, in a typical robotic setting it can be easier to gather some bounding box annotations, for instance by relying on spatial or temporal contextual information. In this perspective, in [29] a visual tracking algorithm was used to automatically generate, in a self-supervised fashion, the sufficient ground truth (bounding boxes and labels) to learn representations from thousands unlabeled videos. One of the problems of self-supervised pipelines, that generate a pseudo ground truth by relying on the predictions of a previously trained detection model, is model drift and degradation.

Another approach to address a weakly supervised scenario is active learning (AL) [30], [31]. In this case, the effort is focused on defining a sample selection strategy, i.e., a policy to choose the most informative samples to be asked for annotation to an oracle (e.g., a human). In [32] the authors proposed to refine object detectors by actively requesting crowd-sourced image annotations from the web, while in [33] a method that combines AL and semi-supervised learning is proposed to improve object detection performance by leveraging the concept of diversity for the active learning policy. While not suffering from model degradation, these methods still require some human effort –even if significantly lower than a full dataset annotation.

In the proposed pipeline, we consider the *Self-Supervised Sample Mining* method [3], [4] (SSM), a weakly supervised approach, which combines (i) a SS technique to generate pseudo ground truth, with (ii) an AL strategy to select the hardest unlabeled images to be requested for annotation. The SSM method was proposed as an end-to-end deep architecture, where the AL and SS processes alternated with the fine-tuning of a Region-FCN (Fully Convolutional Network) [19]. In this contribution, we isolate the AL and SS processes from the Region-FCN and show a simple approach to use them within our fast, on-line learning pipeline [2]. We also opted for training a new model at every adaptation iteration (rather than fine-tuning or modifying the previous one), which was only feasible due to the training speed of our detection method.

III. METHODS

In the scenario considered in this work, a robot is asked to detect a set of object instances in an unconstrained environment (hereinafter referred to as TARGET-TASK).

We assume that the detection system is initialized with a set of convolutional weights, previously trained off-line on a separate set of objects, using the method described in [18]. A first detection model is trained during a brief interaction with a human, in a constrained scenario (the TARGET-TASK-LABELED). The robot then explores the environment autonomously, acquiring a stream of images in

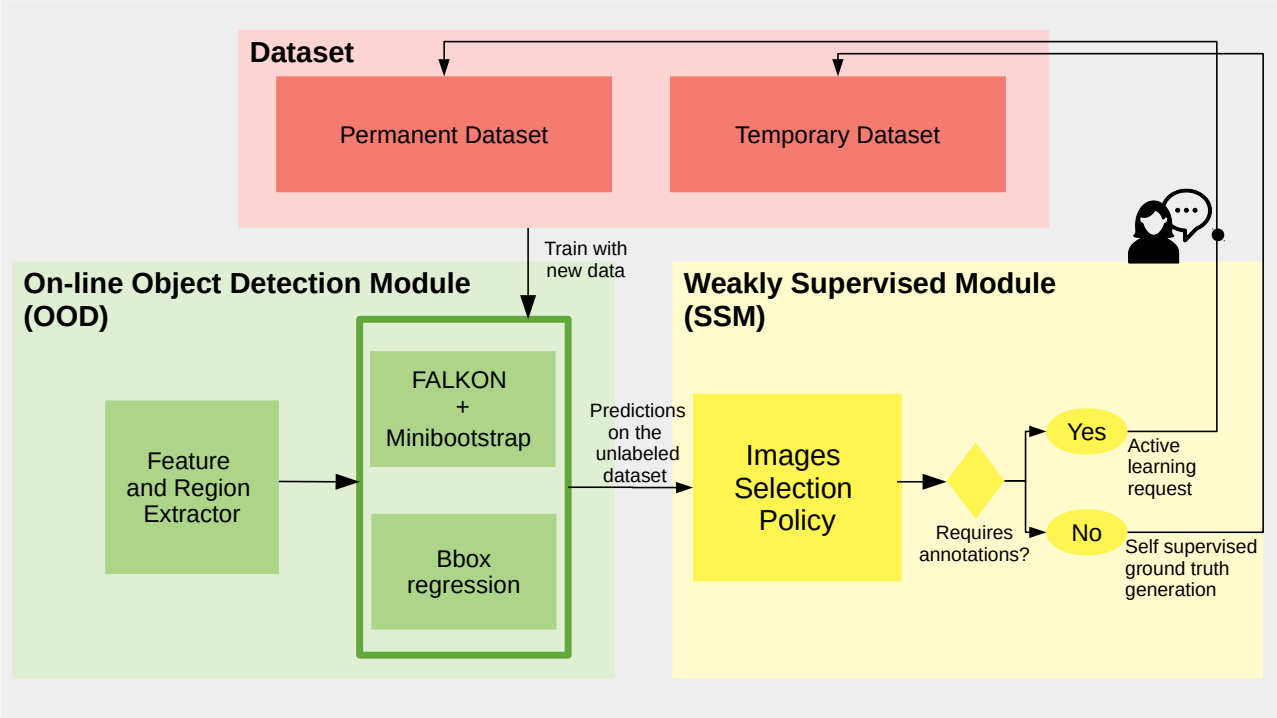

Fig. 1: Overview of the proposed pipeline. The on-line detection system proposed in [2] (green block) is integrated with a weakly supervised method [3] (yellow block) that combines a self supervised technique to generate pseudo ground truth (*Temporary Dataset*), with an active learning strategy to select the hardest unlabeled images to be asked for annotation and added to a growing database (*Permanent Dataset*). We refer the reader to Sec. III for further details.

a new setting. These images are not labeled (TARGET-TASK-UNLABELED) and are used to adapt the detector.

The pipeline uses the the on-line detection algorithm proposed in [2] and an adaptation of the weakly supervised approach of SSM [3], [4]. The detector is adapted thanks to the additional training data which is either automatically labeled by the robot (we call it pseudo ground truth) or labeled with human supervision.

A. Pipeline Description

The proposed pipeline is divided into two main modules (see Fig. 1): an (i) *On-line Object Detection Module* (OOD) and a (ii) *Weakly Supervised Module* (SSM). The first one predicts bounding boxes and labels, and can be trained in few seconds as a new dataset is available, while the second one processes the predictions generated by the former one on a stream of (unlabeled) images in order to generate their annotations.

On-line Object Detection Module. For this first module (green block in Fig. 1) we rely on the method proposed in [2]. This method consists of a (i) first stage of region proposals and feature extraction and (ii) a second stage of region classification and bounding box refinement.

The first stage relies on layers from the Faster R-CNN architecture [18], specifically the convolutional layers, the Region Proposal Network (RPN) [18] and the RoI pooling Layer [17]. In particular, this part is used to extract a number of Regions of Interest (RoIs) from an image and encode them into a set of features. In this work, we considered ResNet50 [34] as the CNN backbone for Faster R-CNN.

The second part is composed of a set of FALKON [25] binary classifiers (one for each class of the TARGET-TASK) and Regularized Least Squares (RLS), respectively for the classification and the refinement of the RoIs proposed at the previous stage. Specifically, the training of the classifiers applies an approximated bootstrapping approach, called Minibootstrap [2]. This approach is used to overcome the well-known problem in object detection of the background-foreground class imbalance, while maintaining a learning time of the order of seconds. Please, refer to [2], for further details about this algorithm.

Weakly Supervised Module. The aim of this module (yellow block in Fig. 1) is to generate a new training set by combining images annotated by the robot autonomously, with those annotated with human supervision. This is achieved with an iterative process [3]. For each iteration, the predictions of the current detection model on the images acquired by the robot (the TARGET-TASK-UNLABELED) are evaluated in order to identify (i) those detections that can be used as training set (pseudo ground truth) and (ii) those that need to be labeled with a human intervention. The dataset resulting from this process is used to train a refined version

of the model with the *On-line Object Detection Module*.

For this module, we rely on the weakly supervised approach proposed in [3]. It combines a self supervision based on a *Cross Image Validation* to select a reliable pseudo ground truth, with an active learning policy to pick the most informative unlabeled samples and ask for their annotation. Specifically, the *Cross Image Validation* is performed for each unlabeled image of the TARGET-TASK-UNLABELED and is designed as follows: the current detection model is tested on an unlabeled image, then, the consistency of the predicted detections is evaluated by (i) pasting them into different annotated images and (ii) using the current detection model to predict them. If the detection is confirmed for the majority of the cases, it is considered consistent (the reliability is measured by a *Consistency score*), and thus usable as pseudo ground truth.

Instead, for the active learning process, the selection criteria is based on the classical uncertainty-based strategy [35] where the policy is to ask for annotations of the least confident samples, (the *Consistency score* computed previously is used as measure of confidence of the image).

B. Training the Pipeline

The learning process of the proposed method is divided into two phases: (i) a fully supervised learning stage with a few seconds of interaction with a human, on the TARGET-TASK-LABELED, in order to get a first detection model, and (ii) a weakly supervised learning stage, where the previously trained detector is used to generate pseudo ground truth, or queries for image annotations, on the TARGET-TASK-UNLABELED.

Fully Supervised Phase. The features provided by the *Feature and Region Extractor* (see Fig. 1) are used as training examples for the FALKON classifiers and the RLS regressors, for region proposals classification and refinement, respectively. For the RLS regressors, we used the method of Region-CNN [16], keeping the same learning objective and loss function. For the classification, we consider a one-vs-all approach (so that a multi-class problem is addressed with a collection of n binary classifiers, where n is the number of classes). For each class, the training set is collected by selecting and labeling region proposals as either positive examples (i.e., belonging to the class) or negative ones (i.e., belonging to the background). The resulting dataset is used to train a binary classifier and it is usually large and strongly unbalanced, due to the fact that the majority of the regions typically depicts background areas. The large size and imbalance of this dataset is addressed by the Minibootstrap procedure [2], which is an approximation of the Hard Negatives Mining procedure adopted in Region-CNN [16] and in [23].

The combination of FALKON, the Minibootstrap and the RLS regressors is used to train a detector on the TARGET-TASK-LABELED. This model will be, consequently, used as a seed model for the weakly supervised learning phase on the TARGET-TASK-UNLABELED.

Weakly Supervised Phase. After the first supervised learning phase, the weakly supervision process on the TARGET-TASK-UNLABELED starts. For this phase we rely on the protocol proposed in [3]. Specifically, this is a process that iterates on the TARGET-TASK-UNLABELED to progressively refine the detection model. Each iteration is structured as follows: the images of the unlabeled dataset are predicted with the current model and the consistency of the predictions is evaluated with the *Image Cross Validation* procedure illustrated above. The images with a high *Consistency score* are added as pseudo ground truth while the ones with a low *Consistency score* or the ones ambiguous for the detector (specifically, the images where the same region is predicted with two positive categories) are added to the set that needs to be asked for labeling.

The dataset composition at each iteration is controlled by a parameter that limits the number of images to be added to both sets, which is defined as a percentage of the TARGET-TASK-LABELED. The strategy adopted to set this parameter in [3] is to allow, for early iterations, a higher number of images to be labeled, while, in subsequent iterations, an increasing number of pseudo labeled images can be added.

After this pruning, the images considered as pseudo ground truth are added to a *Temporary Dataset*, while the ones that need annotation are asked to be labeled and then added to a *Permanent Dataset* (see Fig. 1). Note that, while at the beginning of this iterative procedure the first one is empty, the latter one already contains the TARGET-TASK-LABELED. At the end of each iteration, while the *Permanent Dataset* is retained (it thus grows at each iteration), the *Temporary Dataset* is cleaned. For further details on this weakly supervised approach we refer the reader to [3].

Note that we adopted the protocol of [3], but we replaced the fine-tuning of Region-FCN [19] with the fast learning method proposed in [2], thus reducing the training time at each iteration from minutes/hours to a few seconds, allowing to use the pipeline in an on-line scenario. Another important distinction with respect to the original SSM algorithm is that, in our pipeline, at each iteration the detector is trained from scratch on the composed image set, while in SSM the Region-FCN is fine-tuned with a warm restart from the weights obtained at the previous iteration.

IV. EXPERIMENTS

In this section we first describe the datasets used for evaluation (Sec. IV-A), then we provide details about the setup used for the experiments (Sec. IV-B) and finally we present the performance achieved by the proposed pipeline on two different scenarios (Sec. IV-C and Sec. IV-D).

A. Datasets description

In this section we describe the datasets used for the experimental analysis of this work.

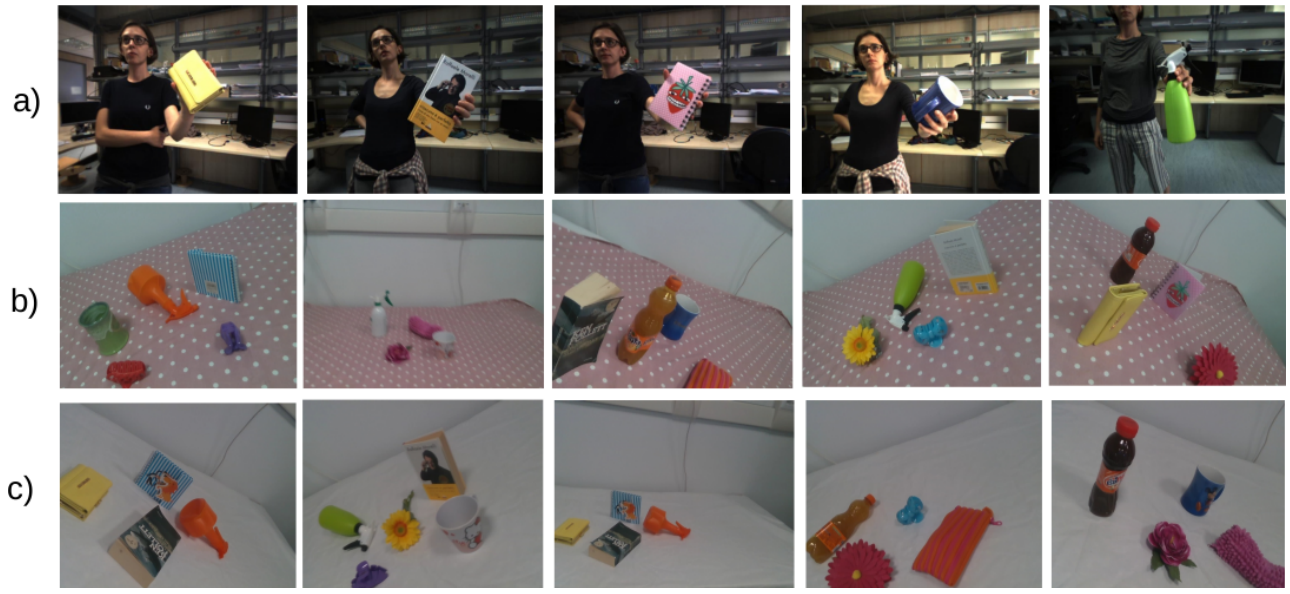

Fig. 2: Examples images of the datasets used for this work: **a)** ICWT dataset; **b)** POIS cloth in the table top dataset; **c)** WHITE cloth in the table top dataset.

iCubWorld Transformations Dataset. The ICUBWORLD TRANSFORMATIONS dataset[2] [6] (hereinafter referred to as ICWT) contains images for 200 objects instances belonging to 20 different categories (10 instances for each category). Each object instance is acquired in two separate days and, for each day, different sequences representing specific viewpoint transformations are collected: planar 2D rotation (2D ROT), generic rotation (3D ROT), translation with changing background (TRANSL), scaling (SCALE) and, finally, a sequence that contains all transformations randomly combined (MIX). The sequences have been acquired with the iCub humanoid robot [7], with an automatic annotation procedure that relies on human interaction in a student-teacher fashion [6]. See Fig. 2 (first row) for some example images.

Table Top Dataset. To prove the generalization capabilities of the proposed integration to different settings, we collected a table top dataset (that will be made publicly available at the same ICWT website) by using the R1 robot [8]. For this dataset we selected 21 objects from ICWT.

The data acquired is split in 2 sets of sequences. In each set we considered a different table cloth: (i) pink/white pois (hereinafter referred to as POIS) and (ii) white (hereinafter referred to as WHITE). For each set we split the 21 objects in 5 groups, and we acquire 2 sequences for each group for the WHITE set, and 1 sequence for each group for the POIS set, gathering a total of 2K images for the WHITE set and 1K images for the POIS set.

For each sequence, the robot is placed in front of the objects and executes a set of pre-scripted exploratory movements to acquire images depicting the objects from different perspectives, scales, and viewpoints. We used a table top segmentation procedure to gather the ground truth of the object locations and labels, and we manually refined them using the *labelImg* tool[3]. See Fig. 2 (second and third rows) for some example images.

B. Experimental Setup

To show the effectiveness of the proposed integration we present results on two different experiments. We firstly validate the pipeline on ICWT, then we consider the scenario of a robot trained with human interaction to detect a set of objects, which needs to adapt and refine the detection model in order to generalize to a different setting. Specifically, in this work we consider as a new setting, the table top dataset described above. This is a challenging task as the robot is trained by a human demonstrator while holding the objects in the hand and it is later required to detect objects when they are placed on a table (see Fig. 2 to compare the two settings). Fast adaptation is required to avoid large performance drop as demonstrated by our experiment.

Note that, when considering the TARGET-TASK-UNLABELED, we simulate the human intervention for providing annotations, by fetching the actual ground truth from the dataset. We report performance in terms of mAP (mean Average Precision) at the IoU (Intersection over Union) threshold set to 0.5, as defined for Pascal VOC 2007 [36].

All experiments reported in this paper have been performed on a machine equipped with Intel(R) Xeon(R) E5-2690 v4 CPUs @2.60GHz, and a single NVIDIA(R) Tesla

[2]https://robotology.github.io/iCubWorld/#icubworld-transformations-modal/

[3]https://github.com/tzutalin/labelImg

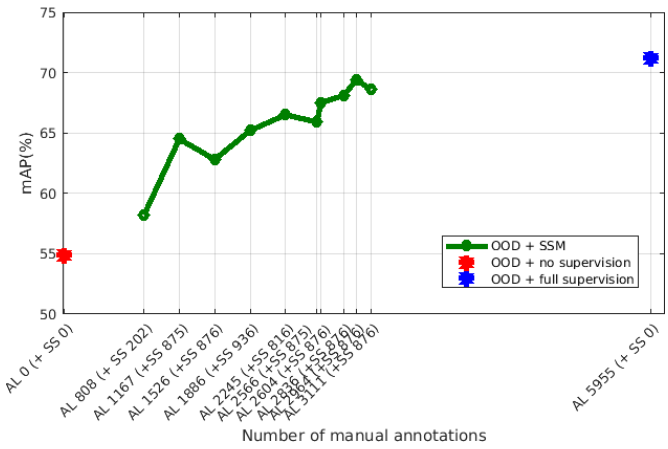

Fig. 3: Benchmark on ICWT. The figure shows (i) the mAP trend of the proposed pipeline, as the number of annotations required on the TARGET-TASK-UNLABELED grows (*OOD + SSM*), compared to (ii) the accuracy of a model trained only on the TARGET-TASK-LABELED (*OOD + no supervision*) and to (iii) the mAP of a model trained with full supervision on the TARGET-TASK-UNLABELED (*OOD + full supervision*). The number in parenthesis represents the number of images selected by the self supervision process at each iteration.

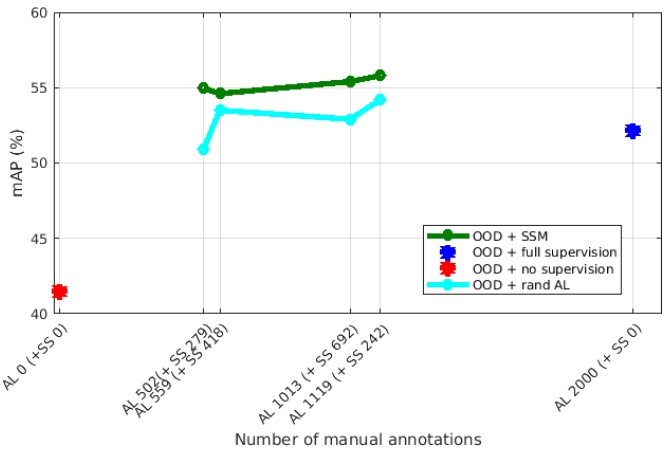

Fig. 4: Benchmark on the table top dataset. The figure shows (i) the mAP trend of the proposed pipeline, as the number of annotations required grows (*OOD + SSM*), compared to (ii) the mAP of a model trained on the TARGET-TASK-LABELED (*OOD + no supervision*) and to (iii) the mAP of a model trained with full supervision on the TARGET-TASK-UNLABELED (*OOD + full supervision*). In this experiment we also compare with the mAP of a model trained only on annotated images randomly selected (*OOD + rand AL*). The number in parenthesis represents the number of images selected by the self supervision process at each iteration.

P100 GPU. Furthermore, we limit the RAM usage of FALKON to at most 10GB.

C. Experiments on the iCubWorld Transformations Dataset

For this experiment, we define as TARGET-TASK a 30-object identification task, considering 3 instances for each 10 categories in ICWT remaining after excluding those used for initializing the CNN backbone. For each object, we then use the TRANSL sequence (for a total of ∼2K images) as TARGET-TASK-LABELED and the union of the 2D ROT, 3D ROT and SCALE sequences (for a total of ∼6K images) as the TARGET-TASK-UNLABELED. This simulates a situation where only a simple sequence is fully annotated and other sequences are not. As a test set, we used 150 images from the MIX sequence of each object, whose annotations have been manually refined adopting the *labelImg* tool[4].

In Fig. 3 we report the mAP trend (**green line**) with respect to the total number of images asked for annotation in the TARGET-TASK-UNLABELED (in parenthesis we specify the number of samples selected by the self supervision process). Note that, as the images get accumulated at every iteration, in order to calculate how many images are required by the robot, one has to take the difference of the indicated number with the one at the previous iteration.

The **red point** shows the mAP on the considered test set, achieved after the supervised learning phase, i.e., after training the detection module on the TARGET-TASK-LABELED. Thus, we consider it our lower-bound. The **blue point** represents the mAP achieved by training the detection module on the union set of the TARGET-TASK-LABELED and TARGET-TASK-UNLABELED (fully manually annotated). Thus, we consider it as the upper-bound of this experiment. As it can be observed, nearly half of the images of TARGET-TASK-UNLABELED are enough to obtain ∼70% of mAP with a drop in performance of ∼1.2% with respect to the fully supervised case.

Each point of the green line has been obtained by re-training a new set of 30 FALKON classifiers, with the Minibootstrap, on the data accumulated after the weakly supervised iteration. As the dataset increases, the training time increments from ∼40 seconds to ∼60 seconds, with an average of ∼55 seconds for each step.

D. Experiments on Table Top Scenario

For this experiment, we define as TARGET-TASK an identification task among 21 object instances chosen from the ICWT –excluding those used to inizialize the CNN backbone. As TARGET-TASK-LABELED, we select a subset of the available images from the TRANSL, 2D ROT, 3D ROT and SCALE sequences (for a total of ∼ 5600 images), while we consider the 2K images of the WHITE table top set (see Sec. IV-A) as TARGET-TASK-UNLABELED and the POIS table top set as test set.

In Fig. 4, we show the result of this experiment. As before, with the **green line** we report the mAP with respect to the increasing number of images asked for annotation, and

[4]https://github.com/tzutalin/labelImg

indicated in parenthesis the number of self-annotated images at each iteration.

Similarly, the **red point** shows the mAP on the considered test set, achieved after the supervised learning phase on the TARGET-TASK-LABELED, while the **blue point** represents the mAP obtained by training the on-line detection module on the union set of the TARGET-TASK-LABELED and TARGET-TASK-UNLABELED (fully annotated).

As it can be observed, just a quarter of the full TARGET-TASK-UNLABELED dataset was enough to train a model with even a higher accuracy ($\sim$55%) than the one obtained with full supervision ($\sim$52%). This may be due to the fact that, by using all images from the TARGET-TASK-UNLABELED, the model may overfit the scenario of the white table cloth, which causes a poorer performance when testing on images depicting a different table cloth. Our findings suggest that AL algorithms may help reducing overfitting, confirming what has been previously reported in the literature (see, e.g., [37]).

One may argue that, in order to avoid the overfitting caused by considering all the images in the TARGET-TASK-UNLABELED (**blue point**), a random sub-sampling of the images to label would suffice. To this end, in Fig. 4 we also compare the proposed approach with a model trained on the same number of images as the ones selected by the AL process, but randomly sampled (**cyan line**). It can be noticed that, while the mAP obtained is relatively high, it also presents a gap with respect to the performance achieved with the integration proposed in this work, demonstrating the effectiveness of the active learning and self supervision processes in choosing the more meaningful samples.

As for the previous experiment, each point of the green line has been obtained by retraining a new set of 21 FALKON classifiers, with the Minibootstrap, on the data accumulated after the weakly supervised iteration. As the dataset increases, the training time increments from $\sim$35 seconds to $\sim$47 seconds, with an average of $\sim$42 seconds for each step.

V. CONCLUSIONS

In this work we proposed a pipeline for on-line adaptation of object detectors in scenarios with limited human supervision. To this end, we extended our on-line detection system from [2] with a weakly supervised method taken from [3]. This latter combines a self-supervision process to generate pseudo ground truth for the most confident predictions, with an active learning strategy to select the hardest images to be asked for annotation. In the integration, we replaced the detection learning adopted in [3] (i.e., the fine-tuning of Region-FCN) with our learning method, which can be trained in much less time (a few seconds), since it relies on the efficient FALKON algorithm [25] and our *Minibootstrap* approximation [2]. Moreover, we show, with the experimental analysis presented in this work, that the effectiveness of the weakly supervised approach of [3] in reducing the annotation effort is preserved.

For this analysis, we simulated the action of asking for human supervision with a process that reads annotations from a database. We now plan to devise an interactive application where the human provides annotations through pointing to objects, and by exploiting spatial and temporal cues to propagate labels in absence of human supervision. This involves the implementation of an active exploration policy that allows the robot to push, pick up and rotate objects to acquire new views, while propagating labels by tracking objects and the strategy proposed in the papers, enriched to actively engage humans when their supervision is required.

From an algorithmic point of view, we plan to study a tighter coupling between the self-supervision and active learning processes, with the *Minibootstrap* happening at each training. In fact, the two procedures both iterate on the dataset in order to extract an effective training set, thus our integration offers an interesting starting point to devise a more efficient and robust sample selection process.

ACKNOWLEDGMENT

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. L. R. acknowledges the financial support of the AFOSR projects FA9550-17-1-0390, BAA-AFRL-AFOSR-2016-0007 (European Office of Aerospace Research and Development), the EU H2020-MSCA-RISE project NoMADS - DLV-777826 and Axpo Italia SpA.

REFERENCES

[1] E. Maiettini, G. Pasquale, L. Rosasco, and L. Natale, "Interactive data collection for deep learning object detectors on humanoid robots," in 2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids), Nov 2017, pp. 862–868.

[2] ——, "Speeding-up object detection training for robotics with falkon," in 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Oct 2018.

[3] K. Wang, X. Yan, D. Zhang, L. Zhang, and L. Lin, "Towards human-machine cooperation: Self-supervised sample mining for object detection," in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, 2018, pp. 1605–1613. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Wang_Towards_Human-Machine_Cooperation_CVPR_2018_paper.html

[4] K. Wang, L. Lin, X. Yan, Z. Chen, D. Zhang, and L. Zhang, "Cost-effective object detection: Active sample mining with switchable selection criteria," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 3, pp. 834–850, March 2019.

[5] 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 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Oct 2016, pp. 4904–4911.

[6] G. Pasquale, C. Ciliberto, F. Odone, L. Rosasco, and L. Natale, "Are we done with object recognition? the icub robots perspective," Robotics and Autonomous Systems, vol. 112, pp. 260 – 281, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0921889018300332

[7] G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, C. von Hofsten, K. Rosander, M. Lopes, J. Santos-Victor, A. Bernardino, and L. Montesano, "The icub humanoid robot: an open-systems platform for research in cognitive development." Neural networks : the official journal of the International Neural Network Society, vol. 23, no. 8-9, pp. 1125–34, 1 2010.

[8] A. Parmiggiani, L. Fiorio, A. Scalzo, A. V. Sureshbabu, M. Randazzo, M. Maggiali, U. Pattacini, H. Lehmann, V. Tikhanoff, D. Domenichelli, A. Cardellino, P. Congiu, A. Pagnin, R. Cingolani, L. Natale, and G. Metta, "The design and validation of the r1 personal humanoid," in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Sep. 2017, pp. 674–680.

[9] A. Zeng, S. Song, K. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, N. Fazeli, F. Alet, N. C. Dafle, R. Holladay, I. Morena, P. Q. Nair, D. Green, I. Taylor, W. Liu, T. Funkhouser, and A. Rodriguez, "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," in 2018 IEEE International Conference on Robotics and Automation (ICRA), May 2018, pp. 1–8.

[10] G. Georgakis, A. Mousavian, A. C. Berg, and J. Kosecka, "Synthesizing training data for object detection in indoor scenes," CoRR, vol. abs/1702.07836, 2017.

[11] M. Schwarz, A. Milan, A. S. Periyasamy, and S. Behnke, "Rgb-d object detection and semantic segmentation for autonomous manipulation in clutter," The International Journal of Robotics Research, vol. 37, no. 4-5, pp. 437–451, 2018. [Online]. Available: https://doi.org/10.1177/0278364917713117

[12] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Sept 2017, pp. 23–30.

[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. E. Reed, "Ssd: Single shot multibox detector." CoRR, vol. abs/1512.02325, 2015.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[15] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," arXiv preprint arXiv:1612.08242, 2016.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[17] R. Girshick, "Fast R-CNN," in Proceedings of the International Conference on Computer Vision (ICCV), 2015.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in Neural Information Processing Systems (NIPS), 2015.

[19] j. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," 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. 379–387.

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

[21] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, 2017, pp. 2999–3007.

[22] K. K. Sung, "Learning and example selection for object and pattern detection," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, 1996, aAI0800657.

[23] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 1627–1645, Sept 2010.

[24] A. Shrivastava, A. Gupta, and R. B. Girshick, "Training region-based object detectors with online hard example mining," in CVPR. IEEE Computer Society, 2016, pp. 761–769.

[25] A. Rudi, L. Carratino, and L. Rosasco, "Falkon: An optimal large scale kernel method," in Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 3888–3898.

[26] D. Dwibedi, I. Misra, and M. Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection," in The IEEE International Conference on Computer Vision (ICCV), Oct 2017.

[27] Y. Zhang, Y. Bai, M. Ding, Y. Li, and B. Ghanem, "W2f: A weakly-supervised to fully-supervised framework for object detection," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[28] P. Tang, X. Wang, A. Wang, Y. Yan, W. Liu, J. Huang, and A. Yuille, "Weakly supervised region proposal network and object detection," in The European Conference on Computer Vision (ECCV), September 2018.

[29] X. Wang and A. Gupta, "Unsupervised learning of visual representations using videos," in The IEEE International Conference on Computer Vision (ICCV), December 2015.

[30] B. Settles, "Active learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2009.

[31] ——, "Active learning," Synthesis Lectures on Artificial Intelligence and Machine Learning, 2012.

[32] S. Vijayanarasimhan and K. Grauman, "Large-scale live active learning: Training object detectors with crawled data and crowds," International Journal of Computer Vision, vol. 108, no. 1, pp. 97–114, May 2014. [Online]. Available: https://doi.org/10.1007/s11263-014-0721-9

[33] P. Kyu Rhee, E. Erdenee, D. K. Shin, M. Ahmed, and S. Jin, "Active and semi-supervised learning for object detection with imperfect data," Cognitive Systems Research, vol. 45, 05 2017.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv preprint arXiv:1512.03385, 2015.

[35] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in SIGIR '94, B. W. Croft and C. J. van Rijsbergen, Eds. London: Springer London, 1994, pp. 3–12.

[36] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," International Journal of Computer Vision, vol. 88, no. 2, pp. 303–338, June 2010.

[37] R. Burbidge, J. J. Rowland, and R. D. King, "Active learning for regression based on query by committee," in Intelligent Data Engineering and Automated Learning - IDEAL 2007, H. Yin, P. Tino, E. Corchado, W. Byrne, and X. Yao, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 209–218.