2017 IEEE-RAS 17th International Conference on
Humanoid Robotics (Humanoids)
Birmingham, UK, November 15-17, 2017

Interactive Data Collection for Deep Learning Object Detectors on Humanoid Robots

Elisa Maiettini[1,2,3], Giulia Pasquale[1,2], Lorenzo Rosasco[1,2,3] and Lorenzo Natale[1]

***Abstract*— Deep Learning (DL) methods are notoriously data hungry. Their adoption in robotics is challenging due to the cost associated with data acquisition and labeling. In this paper we focus on the problem of object detection, i.e. the simultaneous localization and recognition of objects in the scene, for which various DL architectures have been proposed in the literature. We propose to use an automatic annotation procedure, which leverages on human-robot interaction and depth-based segmentation, for the acquisition and labeling of training examples. We fine-tune the Faster R-CNN [36] network with these data acquired by the robot autonomously. We measure the performance on the same dataset and investigate the generalization abilities of the network on different settings and in absence of explicit segmentation, showing good detection performance. Experiments on the iCub humanoid robot [25] show that the proposed strategy is effective and can be used to deploy deep object detection algorithms on a robot.**

I. Introduction

The ability to localize and recognize objects in the scene is crucial for autonomous robots to act in unconstrained environments [20]. Detecting objects from the visual input is essential, and often it represents the first step in the interaction of the robot with the environment.

While planning actions requires full pose estimation in 6D, a common approach adopted in the literature is to split the problem in different stages. The first step is the localization of objects in the image plane. In Computer Vision this task is called 2D object detection and consists in predicting the label and location (e.g., in the form of 2D bounding box coordinates) for each object represented in the image (see, e.g., [37]). In this work, we focus on this task, as a prerequisite for more complex operations in scene understanding.

We consider recent Deep Learning (DL) methods [21], [36], [39], motivated by the remarkable performance they obtain on difficult tasks such as the ImageNet Large-Scale Visual Recognition [37], the MS COCO [23] and Pascal VOC [12] challenges. One of the problems in the adoption of these methods in robotics, is that they require a large dataset of images carefully annotated. Image annotation is particularly demanding for training object detection systems, because this process requires not only object labels but also bounding boxes. In addition, it implies an off-line process, which is unfeasbile for a system that learns on-line.

[1] iCub Facility, Istituto Italiano di Tecnologia, Genoa, IT
[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, IT

Fig. 1: The iCub robot observes some of the objects learned using the proposed pipeline and detects them on a shelf. A video showing the system running has been made available as Supplementary Material.

We propose an approach to train DL methods for object detection that overcomes the lack of manual annotations by exploiting the interaction with a human teacher. Learning happens in a natural, semi-controlled setting, in which objects are presented to the robot by the teacher and the problem of figure-ground segmentation is greatly simplified. This acquisition procedure was validated in our previous work [28], where we showed how this approach can be adopted to acquire large-scale annotated image datasets (i.e. the iCubWorld Transformations[1]). Some example images in the dataset are represented in Fig. 2; annotations are in terms of the label of the object and a surrounding bounding box computed with the depth estimation and segmentation procedure presented in [29].

The contribution of this work is to assess whether this approach can be adopted to effectively train deep object detectors as the recent Faster Region-CNN [36]. More specifically, we first investigate whether such models can learn to predict accurate bounding boxes from imperfect ground-truth. In fact, there are clearly multiple sources of noise in the automatic annotation procedure, which negatively affect the quality of supervision with respect to manual annotations.

In addition, our learning scenario is rather constrained: images depicts isolated objects, well separated from the background, and almost always centered in the visual field because the robot is tracking it with the eyes. The next question, is then, to investigate whether the learned detector can generalize to other, less constrained, scenarios, i.e. when multiple objects are present in the visual scene and figure-

[1] https://robotology.github.io/iCubWorld/

Fig. 2: Examples images from iCWT, depicting 20 objects shown to the robot by a human teacher.

ground segmentation with depth cues is more challenging.

In order to quantitatively assess the effectiveness of this procedure, we i) manually annotate a subset of the iCUB-WORLD TRANSFORMATIONS dataset, on which we measure the performance of the detectors trained with automatic annotations, and ii) collect and manually annotate three additional image sequences, representing a subset of the objects in the dataset randomly placed in different indoor settings. Experiments demonstrate that the proposed learning strategy is effective: DL detectors learn to predict bounding boxes which are more accurate than those available in the training set, and, more importantly, they can generalize well to different settings, providing good detection in absence of explicit segmentation.

The paper is organized as follows: Sec. II reviews related work; Sec. III describes the proposed pipeline, focusing in particular on the data acquisition procedure (Sec. III-A and on the deep architecture adopted for our experiments (Sec. III-B). Sec. IV reports on the experimental results obtained in our study and finally Sec. V draws conclusions and outlines future work.

II. RELATED WORK

In this Section we review the latest architectures for object detection and illustrate conventional training strategies. We motivate the choice of the particular detection method we adopted and relate our contribution to other work in the literature that proposes methods to reduce or automatically compute image annotations.

Deep Architectures for Object Detection and training strategies. Deep Learning methods have advanced the state-of-the-art on object detection. All approaches have at their core a deep Convolutional Neural Network (CNN) [7], [21], [46] often indicated as "feature extractor CNN". This network is then integrated into a more complex "meta-architecture" which relies on extracted features for the localization and recognition (i.e. detection) of the objects in the image [17]. A possible approach is to partition the image using a grid and perform detection in parallel on all the areas (e.g., SSD (Single-Shot MultiBox Detector) [24] and YOLO (You Only Look Once) [34], [35]). Another approach is to perform detection only on a set of "candidate" regions selected with a separate process (e.g., Region-CNN (R-CNN) [16] and its optimizations Fast R-CNN [15], Faster R-CNN [36] and Region-FCN [9]). Approaches in the first group are in general faster, because they do not need a per-proposal processing for detection; on the other hand, approaches in the second group proved to be generally more accurate [17]. Among the various solutions, Faster R-CNN [36] seems to be the most suitable for robotics, as it provides good accuracy while preserving efficiency and real-time performance. In this work we employ this method, however it is fair to say that our pipeline is general and could be applied to other models (e.g. SSD).

The parameters of the feature extractor CNN and those of the additional components for the bounding box prediction must be learned from data. These models perform well when trained on large-scale annotated datasets, which include bounding boxes of the objects (e.g. MS COCO [23] and Pascal VOC2007, 2012 [12]). A common approach to reduce the need for the bounding boxes is to first train the feature extractor CNN – which holds the majority of parameters – on a large-scale image classification task (e.g., the ImageNet Large-Scale Visual Recognition Challenge [37]), and then fine-tune it on the target detection task. Strategies have been proposed to speed up the annotation procedure [26], [27]. Still, these have to be performed off-line and still require a considerable manual effort to provide bounding boxes for a sufficiently large number of objects.

Weakly Supervised Approaches. Researchers have proposed alternative methods, which allow training an object detector from images annotated only with object's presence or absence [2]–[4], [8], [11], [19], [38], [40]–[42], [45]. In some cases adoption of deep CNNs has led to remarkable progress [2]–[4], [8], [19], [41], [42], [45]. However, learning from weak supervision is a difficult task, which leads to performance that are much below those that can be obtained with full supervision.

Exploiting the Context: Learning Features by Training on a Pretext Task. For this reason the literature describes methods that explore ways to extract useful forms of supervision from the contextual information available in real world scenarios (e.g., the spatio-temporal coherence on a sequence of images). Among these, one approach is to exploit the natural structure of the visual data by training a

CNN to solve "alternative" visual tasks [1], [18], [30], [32], [33]. These methods demonstrate that the CNN, which is trained with implicit supervision, can provide good features that can be subsequently fine-tuned on supervised detection tasks.

Exploiting the Context for Computing Annotations. In this paper we follow a similar direction, however, we do not focus on the pre-training of the feature extractor CNN (for which we rely on available models, as it will be detailed in Sec. III-B), but rather exploit the contextual information to compute automatic image annotations (in terms of objects' labels and locations) to fine-tune the architecture for detection. We rely on the interaction between the robot and a human teacher. Object labels are obtained through a speech interface, while motion and depth cues allow the robot to segment the objects and automatically assign bounding boxes. We show that training data acquired in this way is sufficiently accurate to train an object detection model. Interestingly, we show that the trained model is able to generalize to novel scenarios, allowing detection of objects that are static and in presence of clutter, for which motion and depth cues would not be sufficient.

To our knowledge, there are very few works which address this problem (e.g., [31]) and this is the first attempt in the robotics literature to implement an autonomous learning system for object detection. While the acquisition application is unchanged with respect to our previous work [28], the task is made more difficult because it now uses bounding boxes computed from depth cues to train a predictor that detect objects in the full image.

III. Methods

The pipeline proposed in this work is an automatic procedure for extracting and labeling example images for training an object detection network. The robot used for our experiments is the iCub humanoid [25]. In the following, we describe the two main steps in the pipeline: i) data acquisition and ii) model training.

A. Data Acquisition Method

For the first step of data acquisition we used the method employed for the iCubWorld Transformations dataset[2] [28] (shortened to iCWT for simplicity in the following). This method exploits depth information and human-robot interaction to collect labeled images. The procedure is the following: the teacher shows the object in front of the cameras of the iCub. A tracking routine [29], uses stereo vision [14], selecting the pixels from the depth map that are closer to the robot, thus segmenting the object from the background. A bounding box is estimated to surround it, and it is stored as annotation jointly with the label of the objects which is provided verbally by the teacher.

B. Faster Region-CNN Architecture

We chose the region-based method Faster R-CNN [36] as a representative architecture among the ones recently proposed in the DL literature for the same task. This model specifically differs from previous similar approaches (R-CNN [16] and Fast R-CNN [15]) because, instead of employing an external method to provide the object detection network with candidate Regions of Interest (RoIs) –as, e.g., the Selective Search algorithm [44]–, it uses a shallow Convolutional Neural Network called Region Proposal Network (RPN). The RPN, sharing the convolutional layers with the detection network, leads to real-time performance at inference time and allows an efficient training procedure.

As feature extraction CNN, we evaluated and compared two models, whose integration in the Faster R-CNN meta-architecture is publicly available[3]:

- the ZF network proposed by [46],
- the VGG_CNN_M_1024 network proposed by [7].

For both networks, we initialized the training process by adopting for the shared convolutional layers the weights trained on the image classification task of the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2012 [10]. We then fine-tuned by following the 4-Steps Alternating Training pipeline proposed by [36]. This pipeline alternates the optimization of the RPN and of the detection network, thus enabling to learn shared features. In the first two steps, the RPN and the detection network are fine-tuned, by initializing the weights of the shared layers with the ImageNet pre-trained model and training the remaining ones from scratch. In the two latter steps, the shared convolutional layers are kept frozen, and the RPN and the detection network are fine-tuned on the detection task. For fine-tuning, we set the number of training epochs to 6 for steps one and three and to 2 for the other two phases. We left all other parameters unchanged with respect to [36], to which we also refer the reader for a more detailed explanation of the architecture.

IV. Experiments

In this Section we report on the experiments that we performed to quantitatively assess the effectiveness of our approach.

As explained in Sec. III-B, we adopt Faster R-CNN [36] with either the ZF or the VGG_CNN_M_1024 feature extractor CNNs. We address an object identification task among 20 objects, by randomly choosing one object instance for each of the 20 categories available in the iCWT dataset presented in Sec. III-A. Figure 2 shows an example image for each selected object. As training set for the considered task, we used the union of the 4 image sequences available in iCWT for each object, corresponding to the 2D ROT,

[2]iCubWorld website: https://robotology.github.io/iCubWorld/

[3]https://github.com/rbgirshick/py-faster-rcnn/tree/master/models/pascal_voc/ZF

	ovenglove	hairclip	hairbrush	book	ringbinder	flower	cellphone	glass	mouse	wallet	bodylotion	pencilcase	soapdispenser	sprayer	sodabottle	mug	sunglasses	perfume	remote	squeezer	mAP
ZF	0.59	0.78	0.71	0.59	0.52	0.86	0.69	0.83	0.92	0.79	0.84	0.41	0.81	0.61	0.79	0.85	0.61	0.75	0.59	0.77	0.72
VGG_CNN_M_1024	0.53	0.76	0.72	0.55	0.51	0.86	0.62	0.84	0.92	0.77	0.83	0.44	0.75	0.63	0.78	0.85	0.64	0.77	0.64	0.79	0.71

TABLE I: AP for each class and mAP (last column) reported by the two Faster-RCNN models considered in this work when tested on the MIX sequences of ICWT. The reference ground truth is the automatic bounding box provided by the depth segmentation routine.

	ZF	VGG_CNN_M_1024
depth's ground truth	0.71	0.69
manual ground truth	0.75	0.73

TABLE II: mAP reported by the two models on a 3K subset of images sampled from the test set of Table I and manually annotated. It can be noticed that the mAP with respect to the manual ground truth is even higher than the one with respect to the depth's automatic ground truth.

3D ROT, BKG and SCALE viewpoint transformations[4]. We considered the two available acquisition days, both for training and testing, and only images from the left camera, overall leading to a training set of ∼27K images.

We evaluated the system on two settings, respectively in Sec. IV-A and IV-B:

1) First, we assessed the effectiveness of the approach in the same HRI setting: a human holds the object to be detected in the hand. For this test, we could use the MIX sequence available in ICWT, leading to a test set of ∼13k images.
2) Then, we evaluated the ability of the detection system to generalize to a different setting. To this end, we acquired and manually annotated three new image sequences, representing the considered 20 objects randomly positioned on the floor, on a table or on a shelf. We made these sequences already available at the same dataset website.

A. *Evaluation I: Object detection in the same setting*

In this experiment we evaluate the detection performance of the network on a setting similar to the one used for training. Because the training data is segmented and labelled automatically by the robot it inevitably contains errors in the bounding box. This evaluation is therefore important to determine to what extent the detection network is robust to the noise in the training data.

To this end, we considered the two ZF and VGG_CNN_M_1024 models, trained to detect 20 objects from ICWT as described in Sec. III and Sec. IV, and started testing them in the same HRI setting, using the MIX sequences of the considered 20 objects. Testing on these sequences is the best way to evaluate the robustness of the predictions, since, we recall, the object is shown naturally to the robot and can appear in any configuration.

[4]In iCWT the objects are acquired in a way that separate different transformations: planar 2D rotation (2D ROT), generic rotation (3D ROT), translation with changing background (BKG) and scale (SCALE) and, finally, a sequence that contains all transformations (MIX)

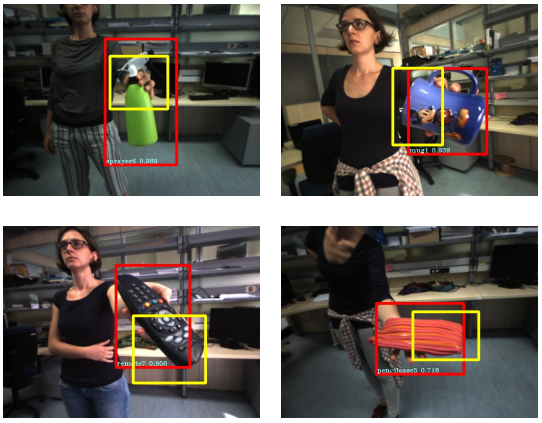
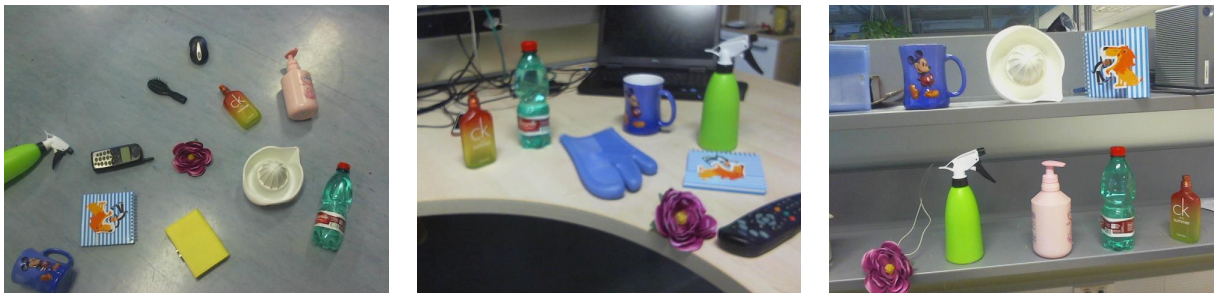


Fig. 3: Example frames where the trained detector (either ZF or VGG_CNN_M_1024) outputs a correct bounding box around the object (red), while the depth segmentation fails (yellow).

In Table I we report, for the two network models, the Average Precision (AP) for each object, with the mean AP (mAP) over all objects (last column). Performance is computed against the ground truth bounding boxes provided by the depth segmentation routine.

It can be first noticed that the reported performance is overall good, in line with the state of the art of deep learning detection systems on other benchmarks (see, e.g., results achieved by Faster RCNN [36] on the Pascal VOC Dataset [12], which consists as well in a 20-class discrimination task). This result is a first important achievement because it shows that these network models, trained with the proposed method, succeeds in localizing and identifying objects in RGB images. Notice that the testing scenario is more challenging, because the object is now localized without the use of depth information. As we will also see in Sec. IV-B, this approach expands the range of possible applications also to settings where the depth cannot be use to localize objects, or not available at all.

In this first test we use as ground-truth the bounding boxes computed automatically using depth information. Because this bounding boxes may contain errors, we further evaluated the system against bounding boxes computed with manual annotation on a subset of the images. We manually annotated this subset of images from the MIX sequences used in the



(a) (b) (c)

Fig. 4: Example images from the three sequences (respectively FLOOR (4a), TABLE (4b) and SHELF (4c)) collected and manually annotated to evaluate the generalization performance of the learned detectors with respect to different indoor settings.

previous test. We adopted the *labelImg* tool[5] and fixed an annotating policy such that an object must be annotated if at least a 50-25% of its total shape is visible (i.e. not cut out from the image or occluded). We annotated 150 frames from each MIX sequence, gathering a test set of 3K images for all the 20 objects that we made available at the ICWT website.

We therefore evaluated the two models on this test set, computing their performance both against the depth's ground truth and the manual ground truth. In fact, a high AP against the depth's ground truth implies that the model learned to predict bounding boxes which are "similar" to the ones provided by the automatic annotation procedure, which, however, may contain noise, be biased or less precise with respect to "ideal" bounding boxes provided by a human supervisor. In Table II, we report the mAP of the predicted bounding boxes against automatic ones (first row) and manually annotated ones (second row). Since the mAP evaluated on the manual ground truth is even higher, it can be inferred that, not only the automatic annotations are sufficient to train good detectors, but these networks also learned to "average out" possible noise in the ground truth, performing thus better than the depth segmentation procedure. In Figure 3 we show some example frames where this effect is evident: while the depth segmentation routine fails to segment the object (yellow), the prediction of the model (red) provides a substantially correct bounding box around it.

B. Evaluation II: Detecting objects in other settings

In the evaluation of the previous Section, training and testing took place in similar settings. We performed a further evaluation to determine to what extent the learning system generalizes to a different scenario. This is important because the training uses images that have (i) the constant presence of a human, holding the object in the hand, possibly generating occlusions, and (ii) the presence of a single object of interest, mostly centered (because the robot was tracking it). In this Section we investigate if this bias [43] affects the generalization properties of the network.

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

We ask therefore whether the networks learned to detect the objects only in these conditions, or are able to generalize to other settings. To this end, we collected and manually annotated three new image sequences, representing three scenes where the objects are randomly positioned respectively on a table, on the floor and on a shelf. These sequences remarkably differ from the ones in ICWT because (i) there is no human presence in the scene and (ii) they contain a variable number of objects, at multiple locations in the image. Finally, also the light and background are different from the ones represented in the dataset. Figure 4 shows an example frame for each sequence (comprising $\sim$ 300 frames):

- The FLOOR sequence depicts 14 out of the 20 objects, lying on the floor.
- The TABLE sequence shows 11 objects on a table with others that are not part of the dataset (like a laptop or a monitor), and hence not to be detected.
- In the SHELF sequence 10 objects are placed on two shelves. The one below is partially shadowed by the one above and, as in the previous sequence, it may contain objects not to be detected.

Table III reports the performance of the two models tested on the three sequences separately, while Fig. 5 shows the predictions of ZF (we obtained similar results with the VGG_CNN_M_1024 model) for two randomly sampled frames from each sequence.

These results show that the performance on these testing sequences remains good, with an average mAPs over the three sequences of 0.58 for the ZF model and 0.52 for VGG_CNN_M_1024. This indicates that the networks succeeded in learning to detect the objects also when these are not hand-held and, as it can be noticed also from the video in the Supplementary Material, which show the predictions of the ZF model running on the iCub while the robot is looking around in these table-top or shelf settings, the proposed approach is a feasible solution to quickly obtain robust and accurate enough object detectors to be used on a general indoor setting.

	ZF	VGG_CNN_M_1024
FLOOR	0.55	0.47
TABLE	0.66	0.32
SHELF	0.53	0.78

TABLE III: Performance (mAP) of the two models considered in this work when tested on the three image sequences described in Se. IV-B.

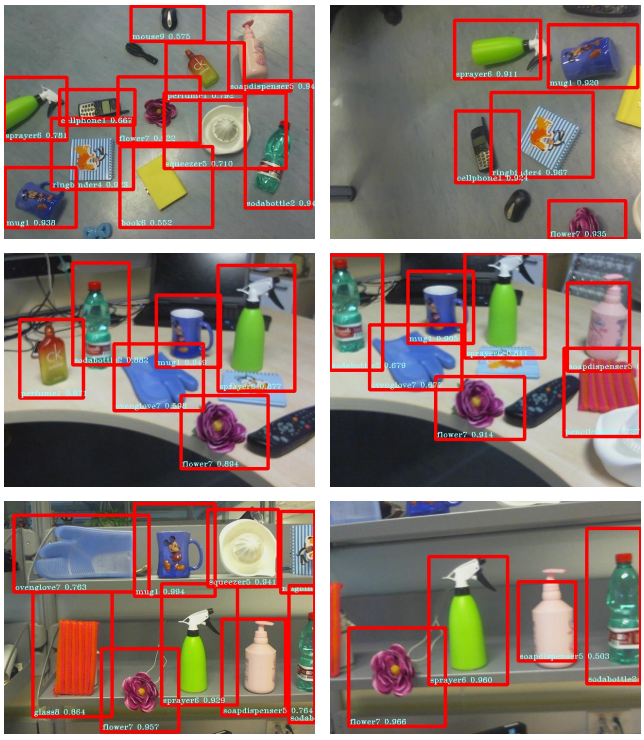


Fig. 5: Example frames, randomly sampled from each sequence, showing the predictions reported by the trained detector (using ZF as feature extractor).

V. Discussion and Conclusions

In this work we propose a procedure to train object detection models in a robotic setting with automatically collected annotated data. The presented pipeline adopts the acquisition setup used for the iCubWorld Transformations dataset [28], which leverages on the interaction of the robot with a human teacher and contextual information on the depth of the scene for segmenting and labeling objects observed by the robot, held by the teacher. We demonstrate that images collected in this way are sufficient for fine-tuning a deep architecture as [36] to successfully detect the learned objects without using any depth information at inference time.

We also observed that the noise in the bounding boxes can be averaged out by the training, and that the learned detector generalizes well to different indoor settings. Overall, our experiments suggest that this strategy can be effectively used to deploy object detection systems to realistic robotic applications.

Current work is focusing on further exploring the capabilities of the models trained with the proposed pipeline, considering other real world settings. We are also working on improving the detection performance, e.g., by exploiting the temporal coherence at test time to refine and stabilize the predicted bounding boxes. In addition we are also assessing the employment of other DL architectures for object detection (e.g., SSD [24]) within the same pipeline.

As future work, we plan to expand the proposed acquisition scenario by including information from the motion of the object (which can be done by incorporating, for instance, the work of [13], [22]). Moreover, human gestures like pointing could also be used in order to focus the robot's attention on objects which are not hand-held. This could be achieved by leveraging on recent improvements in human skeleton detection [6].

Finally, we plan to investigate possible strategies to speedup the current training algorithm in order to provide the robot with the ability of learning on the fly to detect more objects, while interacting with the human. A possible approach which we will evaluate is to start from recent work on incremental object recognition [5] and apply a similar approach to the object detection problem. This would allow to incrementally update the detector, rather than training it from scratch every time new image examples are collected by the robot.

Acknowledgments

The authors would like to thank NVIDIA Corporation for the donation of a Tesla K40 GPU used for this research. This work is funded by the Air Force project FA9550-17-1- 0390 (European Office of Aerospace Research and Development) and by the FIRB project RBFR12M3AC (Italian Ministry of Education, University and Research).

References

[1] P. Agrawal, J. Carreira, and J. Malik. Learning to see by moving. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 37–45, Dec 2015.

[2] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. Weakly supervised detection with posterior regularization. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

[3] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. Weakly supervised object detection with convex clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[4] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[5] R.* Camoriano, G.* Pasquale, C. Ciliberto, L. Natale, L. Rosasco, and G. Metta. Incremental robot learning of new objects with fixed update time. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 2017.

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[7] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

[8] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Multi-fold mil training for weakly supervised object localization. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 2409–2416, Washington, DC, USA, 2014. IEEE Computer Society.

[9] jifeng dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 379–387. Curran Associates, Inc., 2016.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[11] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Localizing objects while learning their appearance. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, pages 452–466, Berlin, Heidelberg, 2010. Springer-Verlag.

[12] 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*, 88(2):303–338, June 2010.

[13] Sean Ryan Fanello, Carlo Ciliberto, Lorenzo Natale, and Giorgio Metta. Weakly supervised strategies for natural object recognition in robotics. *IEEE International Conference on Robotics and Automation*, pages 4223–4229, 5 2013.

[14] Andreas Geiger, Martin Roser, and Raquel Urtasun. Efficient large-scale stereo matching. In *Asian Conference on Computer Vision (ACCV)*, 2010.

[15] Ross Girshick. Fast R-CNN. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015.

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra 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] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. *CoRR*, abs/1611.10012, 2016.

[18] Dinesh Jayaraman and Kristen Grauman. Learning image representations tied to ego-motion. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[19] V. Kantorov, M. Oquab, Cho M., and I. Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *Proc. European Conference on Computer Vision (ECCV), IEEE, 2016*, 2016.

[20] C. C. Kemp, A. Edsinger, and E. Torres-Jara. Challenges for robot manipulation in human environments [grand challenges of robotics]. *IEEE Robotics Automation Magazine*, 14(1):20–29, March 2007.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[22] S. Kumar, F. Odone, N. Noceti, and L. Natale. Object segmentation using independent motion detection. In *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pages 94–100, Nov 2015.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, Zürich, 2014. Oral.

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, and Scott E. Reed. Ssd: Single shot multibox detector. *CoRR*, abs/1512.02325, 2015.

[25] Giorgio Metta, Lorenzo Natale, Francesco Nori, Giulio Sandini, David Vernon, Luciano Fadiga, Claes von Hofsten, Kerstin Rosander, Manuel Lopes, José Santos-Victor, Alexandre Bernardino, and Luis 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*, 23(8-9):1125–34, 1 2010.

[26] Dim P. Papadopoulos, Jasper R. R. Uijlings, Frank Keller, and Vittorio Ferrari. We don't need no bounding-boxes: Training object class detectors using only human verification. *CoRR*, abs/1602.08405, 2016.

[27] Dim P. Papadopoulos, Jasper R. R. Uijlings, Frank Keller, and Vittorio Ferrari. Training object class detectors with click supervision. *CoRR*, abs/1704.06189, 2017.

[28] 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)*, pages 4904–4911, Oct 2016.

[29] Giulia Pasquale, Tanis Mar, Carlo Ciliberto, Lorenzo Rosasco, and Lorenzo Natale. Enabling depth-driven visual attention on the icub humanoid robot: Instructions for use and new perspectives. *Frontiers in Robotics and AI*, 3:35, 2016.

[30] Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *CVPR*, 2017.

[31] Sudeep Pillai and John J. Leonard. Monocular SLAM supported object recognition. *CoRR*, abs/1506.01732, 2015.

[32] L. Pinto and A. Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3406–3413, May 2016.

[33] Lerrel Pinto, Dhiraj Gandhi, Yuanfeng Han, Yong-Lae Park, and Abhinav Gupta. *The Curious Robot: Learning Visual Representations via Physical Interactions*, pages 3–18. Springer International Publishing, Cham, 2016.

[34] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[35] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016.

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Neural Information Processing Systems (NIPS)*, 2015.

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[38] Olga Russakovsky, Yuanqing Lin, Kai Yu, and Li Fei-Fei. Object-centric spatial pooling for image classification. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part II*, ECCV'12, pages 1–15, Berlin, Heidelberg, 2012. Springer-Verlag.

[39] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Robert Fergus, and Yann Lecun. *Overfeat: Integrated recognition, localization and detection using convolutional networks*. 2014.

[40] Parthipan Siva and Tao Xiang. Weakly supervised object detector learning with model drift detection. *2011 International Conference on Computer Vision*, pages 343–350, 2011.

[41] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell. On learning to localize objects with minimal supervision. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1611–1619, Bejing, China, 22–24 Jun 2014. PMLR.

[42] Hyun Oh Song, Yong Jae Lee, Stefanie Jegelka, and Trevor Darrell. Weakly-supervised discovery of visual pattern configurations. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1637–1645. Curran Associates, Inc., 2014.

[43] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1521–1528, 2011.

[44] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[45] Chong Wang, Kaiqi Huang, Weiqiang Ren, Junge Zhang, and Steve Maybank. Large-scale weakly supervised object localization via latent category learning. *IEEE Transactions on Image Processing*, 24(4):1371–1385, 2015.

[46] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013.