Web-Enabled Object Detection with CAT: Unveiling Unknown Instances

Dario Spoljaric
Technical University of Vienna
Vienna, Austria
e11806417@student.tuwien.ac.at

Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Markus Vincze
ACIN - Institute of Automation and Control
Technical University of Vienna
Vienna, Austria
markus.vincze@tuwien.ac.at

***Abstract*—Detecting and recognizing unknown objects in images remains a challenge in computer vision, particularly for service robots. This paper proposes a solution called WILD-CAT (Web-Integrated LoCalization and Detection of Unknown Objects with CAT) that automates the annotation process by leveraging image reverse search engines. The proposed method combines the LoCalization and Identification Cascade Detection Transformer which detects unknown objects, with a web-based image reverse search to annotate unknown objects. Experimental results with never before seen objects validate the effectiveness of WILD-CAT in accurately detecting and labelling unknown objects.**

***Index Terms*—WILD-CAT, OWOD, CAT, web-assisted**

I. Introduction

Detecting and recognising unknown objects in images is a challenge in computer vision, particularly for service robots. To overcome this, an automated solution is required that employs a detector capable of identifying unknown objects and labelling them using internet-derived information. Traditional such as YOLO[5], Single Shot MultiBox Detector (SSD) [2] or R-CNN[1] etc. are trained to detect a fixed number of classes and cannot recognise newly introduced objects, known as the closed-world assumption. However, Open-world object detection (OWOD)[6] relaxes this assumption. In OWOD, a model encounters both known and unknown objects. It must detect and label the known objects while identifying the unknown ones. This process, known as training episodes, involves forwarding unknown objects to an "oracle" for labeling. In the next episode, these labeled instances become the "new known." By continuously expanding its known classes over time without requiring complete retraining, the model can adapt and improve its performance. Detectors like ORE, OW-DETR or CAT attempted to solve the OWOD problem, while still remaining on a human oracle.

The contribution of this work is an advanced method building on the existing detector LoCalization and IdentificAtion Cascade Detection Transformer (CAT)[9] which we call ***Web-Integrated LoCalization and Detection of Unknown Objects with CAT (WILD-CAT)***. WILD-CAT utilises CAT in combination with an online image reverse search to annotate the unknown objects. The reverse search is done using an API for search engines like *google* and *bing / azure*. It is possible to use both engines in parallel and match the labels or using just one. These new labelled results can then be used for visualisation and further training of the CAT detector. Figure 1 shows the overall idea behind the project.

Source code: https://github.com/DarioRepoRuler/wild-cat/tree/main

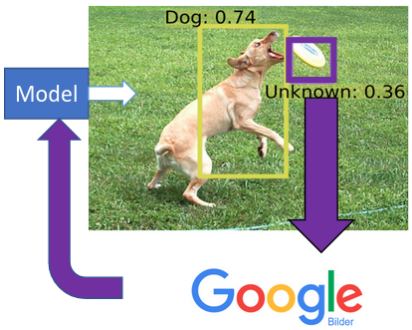

Fig. 1. CAT[9] in combination with an online image reverse search to annotate unknown objects and improve the model

II. Related Work

K. J. Joseph [6] was the first to introduce the **Open World Object Detection** formulation. In the Open World Object Detection setting, the object detection model M_C is trained to detect C known object classes. It can also classify instances of new or unseen classes as *unknown* with a label of zero. Unknown instances U^t can be identified and provided by a human user for n new classes of interest. The model M_C is updated incrementally to include these new classes, generating an updated model M_{C+n} without retraining the entire dataset. The known class set $K_t = 1, 2...C$ is expanded accordingly $K_{t+1} = K_t + \{C+1, ..., C+n\}$. This adaptive learning process continues throughout the object detector's lifespan.

To address this challenge K. J. Joseph also developed the detector **Open World Object Detector (ORE)**[6]. It was the first attempt to solve this problem. The baseline detector of this model is an faster R-CNN(ResNet-50 backbone). ORE consists of two stages first being an class-free(agnostic) region proposal network. The proposals are passed to the second stage named the RoI head, which classifies and adjusts the bounding boxes. The RPN and the ROI-head are adapted to auto-label and identify unknown respectively.

Inspired by the results of Joseph[6] Akshita Gupta approached the OWOD setting with a different detector architecture called **OpenWorld-Detection Transformer (OW-DETR)** [8]. The OW-DETR leverages the performance on unknown object detection by adapting the deformable-DETR[7]. It introduces three key components: an attention-driven pseudo-labeling mechanism, a novelty classification branch, and an objectness branch. The model takes an input image, extracts multi-scale features, and uses a transformer encoder-decoder with deformable attention. Object query embeddings are processed by the branches for bounding box regression, novelty classification, and objectness. The model is trained using dedicated loss terms and can detect both known and unknown objects.
The next improvement in terms of the OWOD setting was made by Shuailei Ma [9] who introduced the **LoCalization and IdentificAtion Cascade Detection Transformer (CAT)** [9] CAT is build on the same architecture as OW-DETR but leverages once again its performance by introducing a new pseudo labelling mechanism for generating robust pseudo labels.

III. Proposed solution

The before mentioned methods must be compared in order to choose the detector with the best performance on detecting unknowns. Therefore the open world evaluation protocol, first introduced in [6] was used. In this protocol the detectors are trained on different amounts of classes in each task. In every task each detector has learned 20 more classes and is evaluated on 80 different classes of the MS-COCO dataset. As can be seen in figure 2 the CAT detector performs best on recognising unknown objects because the *unknown recall* is highest in every task especially in task 3.

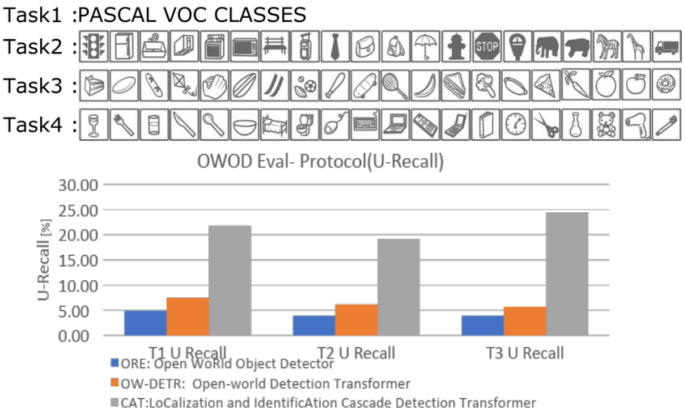

Fig. 2. Unknown recall in OW-evaluation protocol [9]

A. *Overall architecture*

The proposed pipeline includes an pseudo annotator, the CAT detector, an image manager and an API manager. The pseudo annotator is just an additional tool which enables to use custom made pictures without any target data. First the CAT detector is initialised using the pretrained weights from the task 3 of the open world evaluation protocol. The detector is then applied to images yielding to bounding boxes, propabilities and labels. These outputs will be stored in an ".xml" format and the image manager is cropping and collecting the images in a folder. The API manager then takes the images and performes image reverse search using the "*google cloud vision API*[3]" and the "*bing(azure) vision search*" [4]. It then fuses the results to make the result more robust and leverages them to newly annotate the unknown instances in the ".xml" files. Nevertheless it is also possible to use only the results of google or bing with the highest propabilities. These ".xml" files can then be used to train the detector on newly introduced classes which replaces the human oracle in the previously mentioned OWOD setting.

IV. Experiments

The detector in our proposed method is initialized with pretrained weights from task 3, allowing us to determine the known and unknown classes beforehand. Figure 2 provides an overview of the known and unknown classes. We evaluate the detector's performance on the known classes, as shown in Figures 3, 4, 5, and 6. While fruits like apples and oranges as well as microwaves and ovens are occasionally confused, categories such as cars, trucks, and potted plants from the PASCAL VOC dataset are correctly recognized.

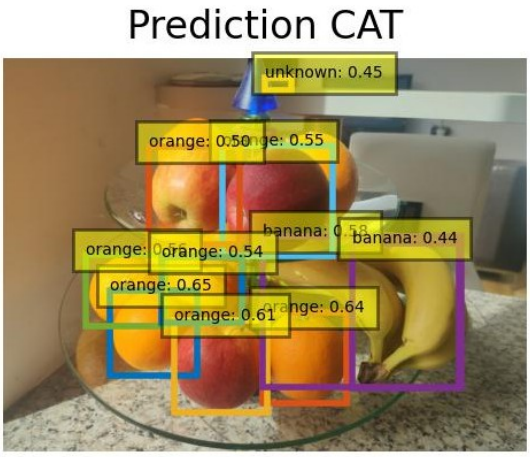

Fig. 3. Known classes: orange, apple, banana

To assess the performance of the detector on unknown classes and evaluate the effectiveness of web-assisted labelling, we conducted experiments on the unknown classes from task 4. As depicted in 7, the detector successfully identified a clock as an unknown object, and the web-assisted labelling correctly labelled it as a clock. Similarly, in 8, the calculator and mouse were recognised as unknown objects, and the web-assisted results provided plausible labels. However, when applied to cutlery 9, the detector exhibited poor performance, only recognising one category, which was subsequently labelled as tableware using web-assistance. The final test on teddy bears 10 yielded similar outcomes.

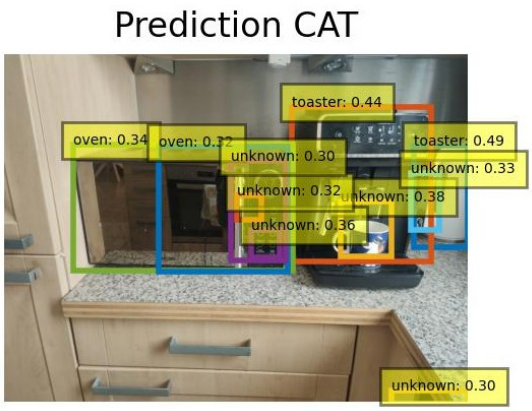

Fig. 4. Known classes: microwave, oven

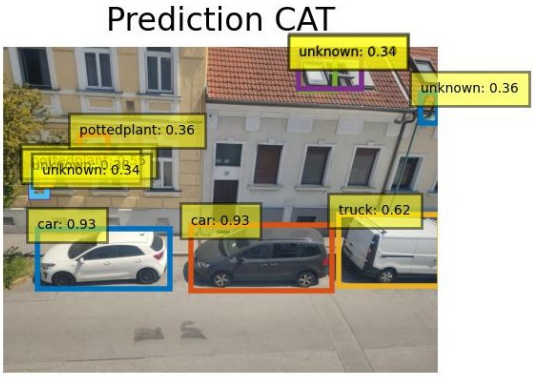

Fig. 5. Known classes: cars, trucks

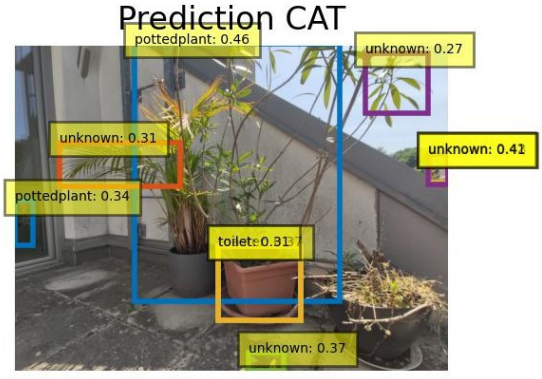

Fig. 6. Known classes: pottedplant

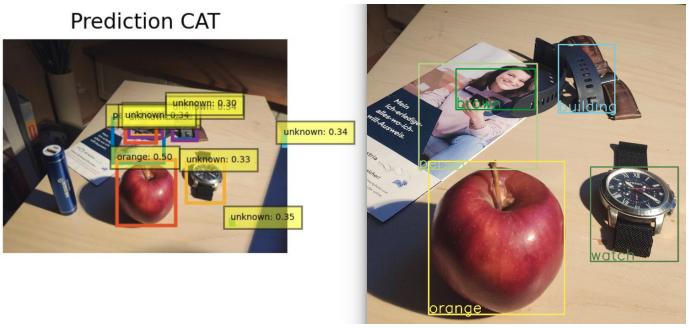

Fig. 7. Unknown detection clock: left CAT results, right web-assisted results

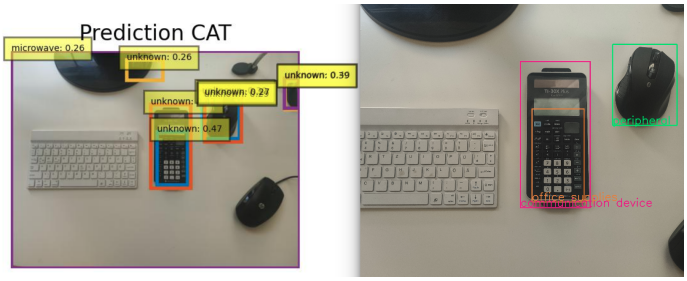

Fig. 8. Unknown detection office supplies: left CAT results, right web-assisted results

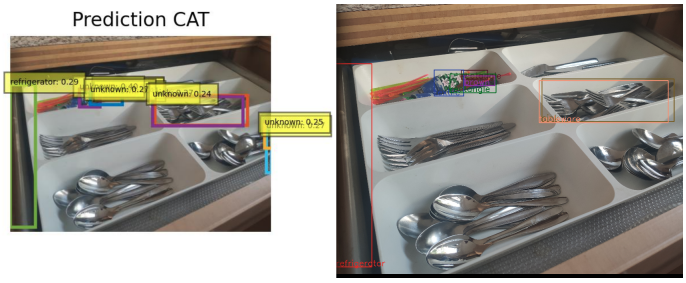

Fig. 9. Unknown detection cutlery: left CAT results, right web-assisted results

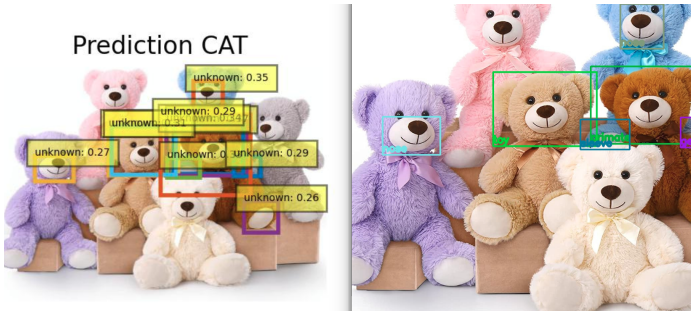

Fig. 10. Unknown detection teddy bears: left CAT results, right web-assisted results

TABLE I
CAT PERFORMANCE COMPARISON

performance metric	CAT (MS-COCO [9])	CAT (selfmade images)
Unknown recall	24.5%	36.58%
mAP	42.3%	48%

V. RESULTS AND DISCUSSION

First the performance of the CAT detector on selfmade images was evaluated and compared with the results of the paper [9]. The results can be seen in table V. One has to note that the dataset of self made images is not as big as the MS-COCO dataset and if the amount of pictures would be increased the results would converge to the ones in [9].

The experimental results validate the effectiveness of our proposed method, WILD-CAT, in detecting and labelling unknown objects with the assistance of web-based resources. However the web-assisted results vary across different search engines and the successful detection of unknown objects heavily relies on the contextual information present in the images. Although the current performance of unknown object detection, with a measured u-recall of 36.58 %, is considered preliminary, it demonstrates promising outcomes even when applied to self-generated images 7 8 9 10. The detector exhibits reliable performance in recognising known object categories, particularly within the PASCAL-VOC data set. However, it faces occasional challenges in distinguishing between similar objects. These findings highlight the potential of WILD-CAT in advancing the field of object detection, and future research can focus on further improving its performance and addressing the identified limitations.

VI. CONCLUSION

The computer vision community has made significant progress in "open-world object detection," but it still relies on human annotators for handling unknown objects. In this study, we propose a novel approach called WILD-CAT (Web-Integrated LoCalization and Detection of Unknown Objects with CAT) that automates this annotation process by utilising image reverse search engines. By incorporating web-assisted labels, the network can be retrained to detect a wider range of known object classes. Although the results are not perfect, they show promising potential. To further improve the method's performance and expand its applicability, future research should focus on retraining the transformer using diverse datasets and integrating additional search engines, such as Pinterest and Alibaba. Additionally, an interesting avenue for future advancements would involve comparing the performance of known classes with well-established detectors like YOLOv5.

REFERENCES

[1] Ross Girshick et al. *Rich feature hierarchies for accurate object detection and semantic segmentation*. 2014. arXiv: 1311.2524 [cs.CV].

[2] Wei Liu et al. "SSD: Single Shot MultiBox Detector". In: *Computer Vision – ECCV 2016*. Springer International Publishing, 2016, pp. 21–37. DOI: 10.1007/978-3-319-46448-0_2. URL: https://doi.org/10.1007%2F978-3-319-46448-0_2.

[3] António Neves and Daniel Lopes. "A practical study about the Google Vision API". In: Oct. 2016.

[4] Houdong Hu et al. *Web-Scale Responsive Visual Search at Bing*. 2018. arXiv: 1802.04914 [cs.CV].

[5] Joseph Redmon and Ali Farhadi. "YOLOv3: An Incremental Improvement". In: *arXiv* (2018).

[6] K J Joseph et al. "Towards Open World Object Detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)*. 2021. arXiv: 2103.02603.

[7] Xizhou Zhu et al. *Deformable DETR: Deformable Transformers for End-to-End Object Detection*. 2021. arXiv: 2010.04159 [cs.CV].

[8] Akshita Gupta et al. "OW-DETR: Open-world Detection Transformer". In: *CVPR*. 2022.

[9] Shuailei Ma et al. "CAT: LoCalization and IdentificAtion Cascade Detection Transformer for Open-World Object Detection". In: *arXiv preprint arXiv:2301.01970* (2023).