Cross Domain Object Detection Benchmark

Kien Ho, Anissa Mokraoui, Pierre Duhamel, Thinh Nguyen

Cross Domain Object Detection Benchmark

1st Kien HO
Université Paris-Saclay
Gif-sur-Yvette, France
kien.ho997@gmail.com

2rd Anissa MOKRAOUI (Advisor)
Sorbonne Paris Nord University
Villetaneuse, France
anissa.mokraoui@univ-paris13.fr

3rd Pierre DUHAMEL(Co-advisor)
CentraleSupélec, Université Paris-Saclay
Gif-sur-Yvette, France
pierre.duhamel@centralesupelec.fr

4th Thinh Nguyen(Co-advisor)
VNU, University of Engineering and Technology
Hanoi, Viet Nam
hongthinh.nguyen@vnu.edu.vn

***Abstract*—Object detection has proven its worth in various real-life tasks, contributing to work efficiency and labor reduction across different domains. However, developing an efficient model often necessitates a large dataset with specific image requirements: 1) diverse viewpoints, settings, and lighting conditions, and 2) coverage of various scenarios. In real-world situations, these prerequisites may pose challenges, especially in domains with limited data availability or when labeling costs are prohibitively high. To address this problem, we propose a cross-domain method utilizing FCOS (Full Convolution One Stage Network). The objective is to leverage a richly labeled dataset (MS-COCO) as the source domain and evaluate FCOS's performance across different target datasets, including DeepFruits, DIOR, Oktoberfest, Clipart1k, IWildCam and ImageNet.**

I. Introduction

Object detection, an essential and challenging task in computer vision, involves a two-step process: localizing instances within an image using bounding boxes and classify them among a predefined set of categories. With the ever-expanding landscape of technology and the emergence of mainstream models like Faster R-CNN [17], YOLO [16], and RetinaNet [15], object detection has undeniably proven its worth in addressing real-world challenges. While object detection models serve as powerful tools, they are not without their limitations and specific requirements. In fact, the effectiveness of an object detector in providing accurate predictions for instance types in a given environment does not ensure comparable accuracy in diverse settings, marked by variations in lighting, viewpoints,and weather conditions. To achieve optimal model efficiency, it usually demands learning from huge dataset with various scenarios to extract general features and make predictions on unseen targets. In real-world scenarios, these limitations may arise in tasks associated with specific domains experiencing data scarcity or when the costs associated with labeling become excessively high.

Cross-domain methods in machine learning is about the creation of new models or modification of optimal architectures to improve the generalization capability across diverse domains or datasets. The goal is to ensure that a well-trained model on a source domain will perform well on a target domain, even with differences between the characteristics of the two domains. This becomes especially critical when there is an abundance of labeled data in one domain but a scarcity of such data in another [14]. Large labeled datasets such as MS-COCO [2], ImageNet [8], or PascalVOC [9] have served as the cornerstone for numerous renowned and state-of-the-art architectures. Notable examples include AlexNet [11], VGG16 [19], ResNet, EfficientNet [20], RetinaNet [14], and Mask R-CNN [8]. The method we propose in this study is using a model pre-trained on large dataset :MS-COCO 2017 [12] as the source domain, then use a transfer learning approach: fine-tuned the model last layer using data of the target domain.

Transfer learning serves as a method to address the problems associated with cross-domain adaptation. Based upon the concept of utilizing knowledge gleaned from a pre-trained model on a sizable labeled dataset (specifically, MS-COCO [12] in our study), the approach seeks to enhance model performance when applied to target domains. Fine-tuning last layer, a specific strategy within transfer learning, refines this process by maintaining fixed param- eters learned from previous layers, while training only the weights of the last layer on a subset of target datasets. This nuanced approach allows the model to tailor its understanding to the unique characteristics of the new target domain [24], facilitating improved performance on specific tasks within that domain and building upon the foundational knowledge acquired during pre-training.

II. Method

A. FCOS Benchmark

In the FCOS study [1], the authors conducted evaluations using the MS-COCO benchmark [2]. Specifically, they utilized the COCO trainval35k split, which comprises 115,000 images, as their training dataset. For validation, a separate minival split from MS-COCO [2], containing 5,000 images, was employed. The final evaluation of their method took place using the COCO testdev-split dataset, consisting of 20,000 images. MS-COCO [2] encompasses a total of 80 categories. In the case of FCOS [1], multiple versions were constructed with modifications in the backbone CNN type, including Resnet-50 [9], Resnet-101 [9], and HRNet [10]. For our comparison, we will focus on FCOS-Resnet50-FPN as it is mentioned as the default setting in the FCOS work [1]. With Resnet-50 as the backbone, FCOS has achieved [1]

AP(average precision) of 37.1%. With some modification in normalization layers, the authors has reached AP of 38.6%.

TABLE I
DOMAIN DATASET INFORMATION

Domain	Dataset	Classes	Bboxes per Image
Aerial	DIOR	20	8.2
Agriculture	DeepFruits	7	5.6
Animal	iWildCam	3	1.5
Cartoon	Clipart	20	3.3
Food and Beverage	Oktoberfest	15	2.4
Common objects	ImageNet	10	1.36

B. Cross-domain benchmark

To assess the adaptability of the FCOS [1] model across diverse domains, we have selected datasets representing a spectrum of object relationships, ranging from close to more distant associations. Specifically, the datasets chosen for this study include: DeepFruits, DIOR, Oktoberfest and Clipart. The difficulty gradient in terms of domain characteristic relationships is defined as follows: DeepFruits, ImageNet, IWildCamp, Oktoberfest, Clipart and the hardest one is DIOR.
This difficulty spectrum is formulated by the contextual scenario of images, image resolution and number of overlapping categories. eg: DeepFruits present little challenges, since its categories overlap with those in the COCO dataset, and they share some contextual background similarities. This commonality with COCO categories and contextual features may facilitate adaptation for the FCOS model.
We first pre-trained FCOS model with a Resnet-50 backbone, FPN(Feature Pyramid Network) and a set of CNN layers in FCOS head with COCO dataset. After that the learnt weights will be transferred to the second phase. This approach ensures that only the last layer learns the generalization specific to the target domain dataset, as illustrated in Figure 1

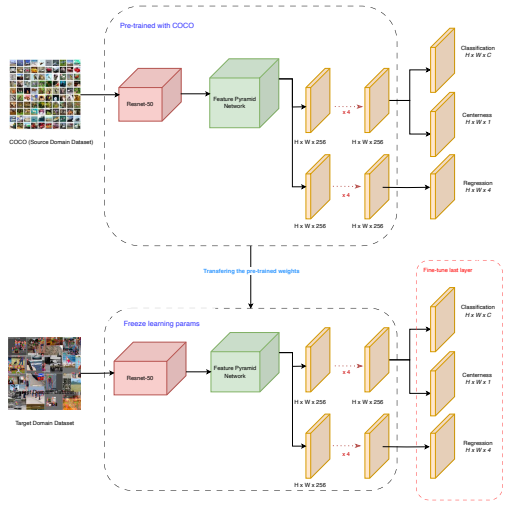

Fig. 1. Fine-tuning process on source and target dataset

C. Few-shot benchmark

In the work by Wang et al. [12], the author proposed a two-stage method: pretraining the model on a data-abundant source dataset and later fine-tuning the last layers. This method has demonstrated impressive results, outperforming several meta-learning-based approaches. With the aim of conducting experiments to assess FCOS's performance in few-shot learning, we adopt a strategy of fine-tuning only the last layer. Two additional datasets, IWildCam [7] and ImageNet [8], will be used, with a small proportion of data sampled to form the training set for performance measurement ImageNet [8] is considered relatively close in terms of object relationships, although it encompasses a significantly larger number of categories, surpassing 10,000 object classes. Since it might be hard to make a statistic measurement for all types of categories in ImageNet [8], we might make a few-shot test to evaluate its adapation with a set of 10 random categories selected.
IWildCam [7] dataset contains 3 categories: animal, vehicle and human. We believe that with animal class the AP metric will be high since with features extracted from the MS-COCO dataset, the FCOS [1] model can adapt well to characteristic of IWildCam [7] dataset. I also make a few-shot test measurement for IWildCam.

III. EXPERIMENTS

A. Implementation Details

As specified, the configuration involves freezing both the Backbone ResNet-50 and FPN layers. This freezing extends to the FCOS head, encompassing the classification and bounding box regression branches, with the exception of the last layer just before the final output, which remains open for fine-tuning. To ensure a fair evaluation, the obtained results will be compared with those of the ResNet50-FPN FCOS architecture. In terms of specifics, the implementation uses an IoU threshold of 0.5, incorporating hyperparameters α set to 0.25 and γ set to 2.0 (2 hyper-params for focal loss function). The initial learning rate of the model is adjusted based on the target datasets, recognizing that fine-tuning FCOS [1] on a new dataset may improve learning efficiency and stabilize parameter updating, particularly when pre-training the model on a smaller dataset. Depending on the dataset, final layer of classifcation branch will be modified adaptively to align with the specific number of categories.

B. Target Datasets Approach

- **DeepFruits (Agriculture).** This dataset is in domain of agriculture which includes images of six different fruits. The dataset provided by author of this paper lacks sufficient data for "Rock melon." To ensure a fair and consistent evaluation, the performance assessment will evaluated based on 5 datasets for: "sweet pepper" (capsicum), apple, avocado, mango, and orange, as data for these fruits is readily available. As a standard practice, we will designate 25% of the images from

TABLE II
CROSS-DOMAIN TASK: MAP SCORES ON DIFFERENT DATASETS

Dataset	mAP	mAP50	mAP75	mAP(small)	mAP(medium)	mAP(large)
DeepFruits	60.33	83.08	62.25	35.75	42.05	63.64
DIOR	17.47	30.24	17.91	1.44	8.21	20.35
Oktoberfest	35.03	52.82	38.61	0.00	10.74	39.33
Clipart1k	37.62	37.61	21.30	11.75	21.79	21.78

TABLE III
FEW SHOT TASK: MAP SCORES ON DIFFERENT DATASETS

Dataset	mAP	mAP50	mAP75	mAP(small)	mAP(medium)	mAP(large)
IWildCam	11.535	21.123	10.555	5.175	7.148	13.892
ImageNet	48.262	66.558	50.281	0.000	23.565	50.318

the training set as the validation set. This split will be utilized to monitor the training/validation loss curve and mitigate the risk of overfitting.

- **DIOR (Aerial Images).** "DIOR dataset includes 23,463 RS (Remote Sensing Data) images, 192,472 object instances, and 20 object categories. The image size is 800 × 800 pixels and the spatial resolution range is from 0.5m to 30m" as stated in the work [4]. We will use 12340 images for traing, 1690 for validation set and 9433 images for testing. We use such an big training dataset due to the substantial difference in characteristics between the target domain and the source domain.
- **Oktoberfest (Food and Beverage).** According to the paper [5] which dataset is provided, the original video dataset captures the scene at a beer tent over eleven days from camera angles. The training dataset includes of 1100 images with instance-level annotations. 85 test images were extracted to assess the models' performance. This test set was setup with challenging scenarios, such as images featuring numerous closely positioned objects, substantial occlusions caused by waiters, and instances of motion blur. I will split training set into 2 splits: 888 images for training and the left 222 images for validation.
- **Clipart1k (Cartoon).** The Clipart1k dataset includes 1000 images collected from the CMPlaces dataset [6], along with images obtained from two image search engines, Openclipart and Pixabay, as mentioned in the study [13]. The dataset is divided into training and testing sets, each containing 500 images. It includes a total of 20 categories, which represent a subset of the 80 categories present in the MS-COCO dataset [2].
- **ImageNet (Common Objects).** The ImageNet dataset [8], comprising millions of images spanning over 20,000 categories, poses a considerable challenge for comprehensive training and testing. To address this, a curated subset featuring 1,000 object classes, encompassing 1,281,167 training images, 50,000 validation images, and 100,000 test images, is provided on the official ImageNet site for research purposes [8]. For our few-shot testing scenario, we randomly select 10 classes from the available 1,000 categories. Subsequently, we create a subset from the original dataset containing instances solely from these 10 selected classes. With a training set comprising 100 images, a validation set of 50 images, and a testing set of 349 images, our objective is to assess the ability of FCOS [1] to generalize learning knowledge effectively, even when provided with a limited amount of training data.
- **IWildCam (Wild Animals).** The IWildCam dataset comprises more than 200.000 images divided in 3 different categories: vehicle, human and animal. Since it is suggested to focus on animal classifcation due to the accurate label and the original task of IWildCam competition [7] is about detecting animals in camera images, we will only focus on the AP of class "animal" in this experiment. For our few-shot testing scenario, we randomly select 300 images for training, 50 images for validation and 2000 images for testing dataset.

CONCLUSION

In the scope of this thesis study, we have made some experiments to measure the performance of FCOS across diverse domains utilizing a transfer learning technique known as last layer fine-tuning. When considering mean Average Precision, the ascending order of FCOS adaptability is observed to be as follows: Aerial Images (DIOR), Cartoon (Clipart), Food and Beverage (Oktoberfest), and Agriculture (DeepFruits). Furthermore, we generated figures to facilitate a comparison between predicting common classes in the source domain and newly introduced classes within the target dataset.

A noteworthy observation is that, with the DeepFruits dataset, FCOS demonstrates effective generalization, performing well even on fruit types like 'mango' and 'avocado.' The accuracy achieved is comparable to or even surpasses that of classes originating from the original source dataset.

Across all tested datasets, there is a consistent under-average mean Average Precision (mAP) for FCOS [1] when detecting small objects. This prompts consideration of alternative strategies, such as image resizing or partitioning into smaller

segments, which could significantly enhance FCOS adaptation across diverse target domains.

REFERENCES

[1] Z. Tian, C. Shen, H. Chen, and T. He. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE International Conference on Computer Vision, pages 9627–9636, 2019.

[2] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. CoRR, abs/1405.0312, 2014.

[3] I. Sa et al. Deepfruits: A fruit detection system using deep neural networks. Sensors (Basel, Switzerland), 16(8):1222, 2016.

[4] Y. Zhan, Z. Xiong, and Y. Yuan. Rsvg: Exploring data and models for visual grounding on remote sensing data. IEEE Transactions on Geoscience and Remote Sensing, 61:1–13, 2023.

[5] A. Ziller, J. Hansjakob, V. Rusinov, D. Zügner, P. Vogel, and S. Günnemann. Oktoberfest food dataset. 2019.

[6] M. V. Conde and K. Turgutlu. Clip-art: Contrastive pre-training for fine-grained art classification. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, June 2021.

[7] S. Beery, A. Agarwal, E. Cole, and V. Birodkar. The iwildcam 2021 competition dataset, 2021.

[8] M. V. Conde and K. Turgutlu. Clip-art: Contrastive pre-training for fine-grained art classification. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, June 2021.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

[10] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao. Deep.

[11] high-resolution representation learning for visual recognition, 2020.

[12] X. Wang, T. E. Huang, T. Darrell, J. E. Gonzalez, and F. Yu. Frustratingly simple few-shot object detection. CoRR, abs/2003.06957, 2020.

[13] . Xu, Y. Sun, Z. Yang, J. Miao, and Y. Yang. H2fa r-cnn: Holistic and hierarchical feature alignment for cross- domain weakly supervised object detection. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 14309–14319, 2022.

[14] M. He, Y. Wang, J. Wu, Y. Wang, H. Li, B. Li, W. Gan, W. Wu, and Y. Qiao. Cross domain object detection by target-perceived dual branch distillation, 2022

[15] Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, number 10, pages 2980–2988, 2017.

[16] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. CoRR, abs/1506.02640, 2015.

[17] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. CoRR, abs/1506.01497, 2015.