https://doi.org/10.2352/EI.2023.35.7.IMAGE-283

Robust Tracking of Industrial Objects across Environments from Small Samples in Single Environments using Chroma-key and Occlusion Augmentations

Yan-Ming Chiou, Bob Price. Palo Alto Research Center – A Xerox Company, Palo Alto CA, 94034, USA

Abstract

Training deep models that can be deployed on embedded systems to robustly detect and track highly specialized industrial objects in a variety of field environments remains very challenging. Large Deep Foundation models (e.g., [yuan21]) make it easier than ever to detect and track everyday objects but do not work as well for specialized industrial objects. These models are often very large and not suitable for deployment on embedded systems. In this work we show that the use of a chroma-key like substitution combined with artificial occlusion generation allows one to capture a small number of images of objects under a fixed background in the lab and then generalize them to novel backgrounds that work in the real world under realistic conditions improving detection of occluded objects by 4% and improving detection in different environments by 44% over state-of-the-art augmentation methods such as MOSAIC.

Introduction

Detection and tracking of objects is key to many applications such as security, robotics, and industrial process monitoring. In recent years large pretrained foundation models [yuan21] have shown good performance when fine-tuned on detection and tracking tasks, but they are typically trained on common classes of everyday objects such as dogs, or cars that may not generalize to industrial parts and can be very large making it difficult to use in embedded systems found in consumer electronics, process monitoring and robotics. Industrial parts often have unique textural and specular properties that make them challenging. More tractable models such as YoloX [Zheng21] can be used on embedded devices when trained with sufficient data, but it can be difficult to acquire images and get them labeled by experts for unique industrial domains. There is an urgent need for new methods for building robust detectors from a small number of samples captured in a limited number of settings that provide tractable but robust detections on real world images.

Related Work

Augmentation is a long-standing technique for improving computer vision models. There are many popular packages implementing standard transforms such as rotation, flipping horizontally or vertically, adding noise of various kinds, changing brightness and contrast, hue shift, gamma transforms, etc. Popular libraries include scikit-image transforms, pytorch vision and Albumentations (https://albumentations.ai/). However, these methods do not address the fact that the background is similar in all of the images making it difficult for the deep learning optimization to reliably characterize what part of the image is the object and what part is irrelevant background.

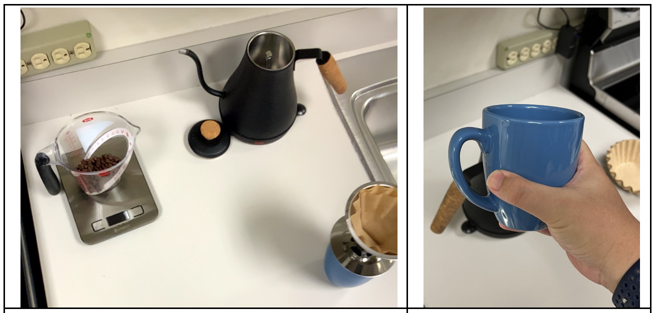
Yolo V4 introduced Cutout augmentations consisting of random boxes drawn over the image [Bochkovskiy20]. Yolo V4 also introduced CutMix or MOSAIC [Bochkovskiy20] which puts together copies of images to generate new images with novel class combinations, crops, rotations and variance in the number of objects. In MIXUP augmentation [Zhang17], objects from different classes are blended over each other to create soft classification problems reducing overfitting. These methods reduce the sensitivity of the detector to the specifics of individual images but still don't focus attention on the difference between foreground and background details in an image.

Challenges and Objective

Our objective is to develop an augmentation pipeline that allows us to capture industrial components during a single lab session and artificially augment these images to train a highly generalizable ego-centric object detector that can be deployed in the field to recognize objects in arbitrary scenarios.

Capturing objects in a lab setting can result in the object detector overfitting to background details. Models trained on individual objects in the lab can also fail in real-world scenarios, where objects do not appear isolated and can be occluded by other objects or by the hands of humans interacting with them.

To protect our proprietary applications, we demonstrate the method on kitchen objects. Figure 1 illustrates two occlusion problems: object-object occlusion, and hand-object occlusion. When we deployed an ego-centric object detector trained on isolated objects against a fixed background to detect the objects in a kitchen, the model could easily fail to detect the objects when occluded or too close to each other (Figure 1a). In the same scenario, when the user picks up the cup object the ego-centric object detector may fail to detect the cup object (Figure 1b).



(a) Object-Object occlusion
(coffee filter occludes cup, measuring cup occludes scale)

(b) Hand-Object occlusion
(Hand occludes cup)

Figure 1 Occlusion of common kitchen objects

Approach

The failures described above arise due to overfitting on the training data. We therefore need to force the network to pay attention to key features in the training images and ignore irrelevant ones. We developed several augmentation methods to enhance the robustness of the detectors. We use background augmentations to improve robustness to new locations and foreground augmentations to improve robustness to occlusion. While these methods reduce the network's ability to use local context clues to identify objects (e.g., the lid of a bottle is suggested by the bottle underneath), these augmentations are critical to make data captured in artificial contexts generalize to the real world. The techniques are explained in the next two sections.

Chroma-Key Background Augmentation

In chroma-key background augmentation, objects from the domain are photographed from various angles against a background with a fixed color and matte finish. We then use chroma key substitution to replace the colored background in these images with various textures. The substitution of a variety of backgrounds decouples the foreground characteristics of the object from the background.

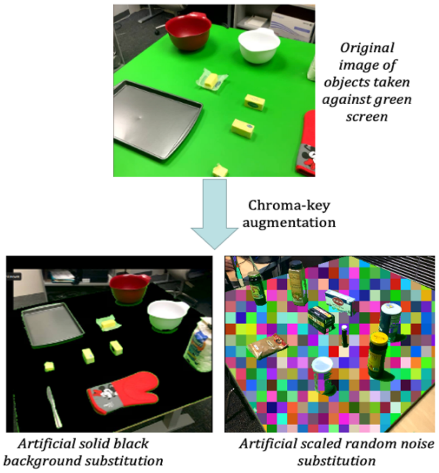

Original image of objects taken against green screen
Chroma-key augmentation
Artificial solid black background substitution
Artificial scaled random noise substitution

Figure 2 Example of background chroma-key augmentation with solid background and random noise

In the most common technique, the image is converted from the RGB color space to the HSV color space to make the hue of each pixel independent of saturation and value. One can then threshold on the hue values to pick out one color from the image. Morphological operators can be used to erode masks to tighten the boundaries. We observed that threshold methods can be problematic when there are shadows and reflective surfaces causing ambiguous regions. Continuous blending based on the closeness of the hue to a reference value resulted in fewer sharp edge artifacts between foreground and background in the presence of shadows and reflections. To avoid problems at the edge of the color space, we used an angular measure between pixels and the reference hue that wraps around the red and violet ends.

We experimented with many types of substitutions, including solid colors, random noise at various scales, and natural backgrounds at multiple scales (Figure 2). White and black were chosen to maximize and minimize energy input to the early stages of the network. Noise was used to prevent overfitting on details around edges of objects.

We found that using natural backgrounds increased performance over artificially generated patterns. However, it was important to crop, scale, translate, and hue transform the background to maximize performance (Figure 3). It is also important that the background image does not contain instances of the foreground objects to be recognized. Having unlabeled target objects in the background of the dataset would cause the network to put these targets into the background class for those images.

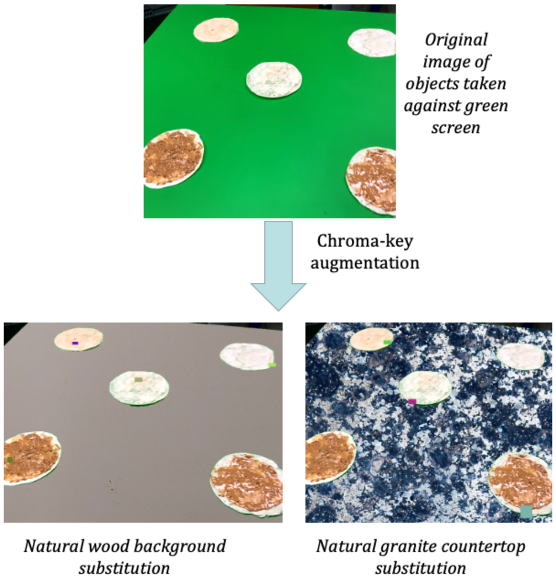

Original image of objects taken against green screen
Chroma-key augmentation
Natural wood background substitution
Natural granite countertop substitution

Figure 3 Example of Chroma-key Augmentation with realistic background

Foreground Augmentations increase robustness to occlusion

Foreground augmentations force the network to develop the ability to recognize an object from a subset of its features. The augmentations we describe here are related to Cutout [DeVries17] in which black boxes are randomly generated over the whole image. In our application we are doing object detection rather than classification of entire images so random boxes could occlude entire objects. We found the method did not work well for our data. We therefore developed three foreground augmentations: *circle grid, curtain* and *object relative random boxes*. Examples of the described augmentation methods are shown in Figure 4.

The *circle grid* augmentation is a simple augmentation which just draws a grid of objects over the image to force the neural network to focus on a portion of the object instead of the whole shape of the object. It relies on the grid objects being small and having a variety of images of each object to ensure all parts of object are visible around the grid objects over the whole dataset. Circle grid augmentations were not as effective as the next two methods.

Curtain augmentations are intended to deal with object-object occlusion where a significant axis of the object is occluded by a neighboring object. A curtain is an occlusion that goes completely across the local bounding box of the object from left-to-right or top-to-bottom. In this method, the curtain is randomly assigned to cover different portions (0%, 25%, 75%), different colors, and different directions (left, right, top or bottom) of the object based on its bounding box information.

Object relative random-box augmentation are intended to make detection robust to hands on the object. It uses the ground truth bounding box information to create small rectangular masks within the objects. This method is the most similar to the CutOut augmentation method [DeVries17], but we randomized the size/fraction of the mask within each object within the image. We also substituted various textures instead of simply blacking out the rectangle. Testing showed that object relative random-box augmentations worked best when allowed to cover up to 75% of image.

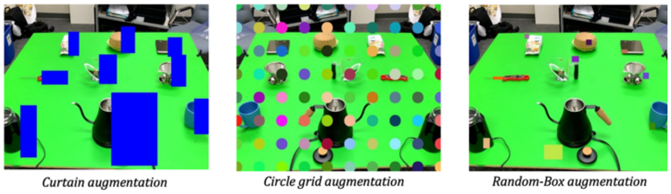


Figure 4 Different type of foreground augmentation methods

The above-described foreground and background augmentations can be randomly combined to generate many different images from a single original (Figure 5 and Figure 6). In the examples shown below, we generate seven different synthetics images from each original.

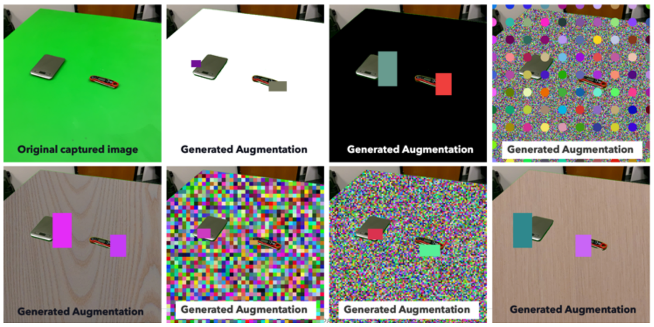


Figure 5 : Example showing a combination of different foreground and background and augmentations to create many images from one original image of a cup and thermometer.

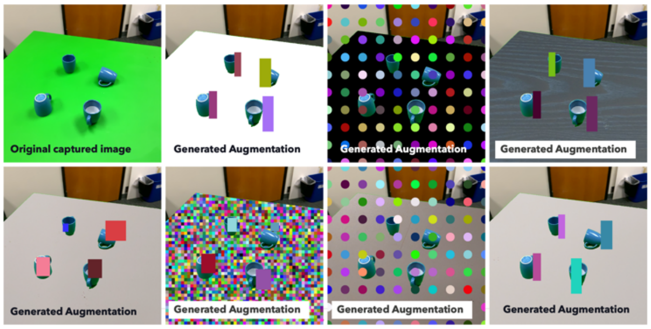


Figure 6 A second example of background and foreground augmentation on variously oriented cups.

Methodology

While this technique is designed for specialized industrial domains, our public results are in the kitchen domain. We build and test a model for detecting various handheld kitchen implements. While public datasets containing kitchen implements exist, we mean to illustrate how one can build a robust detector from images taken in limited contexts rather than advance the detection of kitchen implements. Experimental Setup

We created a colored plywood background using a matte finish Behr "Gamma Sector Green" from the Disney collection and clamped it to a table. We then arranged a subset of objects on this background. Usually 6 to 9 objects at a time depending on the size of the objects. We then used our WARHOL technique [Shreve20] to automatically acquire many images of these objects from multiple angles using continuous video capture and augmented reality anchoring. This particular method of capture is not critical to the results described here but is convenient.

Datasets

We generated four datasets: a baseline dataset taken against a fixed background, an augmented dataset built on the baseline and two test datasets of natural images. The baseline dataset contained 3000 images of 22 classes such as plates, cups, kettles and some foods such as tortillas and peanut butter against a green background.

For the augmented dataset, each image appears in its original form along with 10 augmentations chosen randomly from those described above resulting in an augmented dataset of 33000 images.

For the real-world background test dataset, we collected the same 22 classes on two different scenarios: a) a wooden table environment with different background and lighting conditions and b) a wooden table environment where one or more human hands occludes various objects with different gestures and angles. In Figure 7, the bottom row shows natural images of kitchen implements taken on a wooden table. The top row shows natural images of objects taken on a wooden table that also include hands occluding kitchen towels, plates, cutting boards and tortillas.

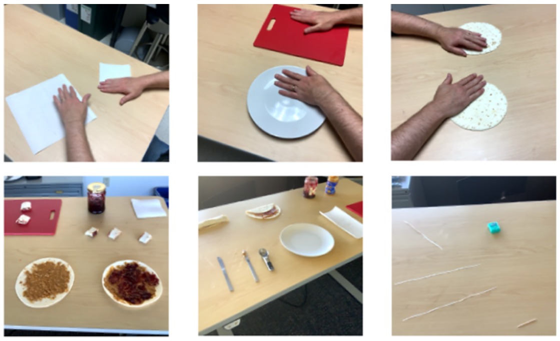


Figure 7 Natural test Images with real world backgrounds (no green tables!) and occlusion by human hands

Training of Models

To create a strong baseline, we train the Megvii implementation of the YOLO X model [Zheng21] with state-of-the-art Mix-up technique [Zhang2017] and the MOSAIC augmentations developed for Yolo V4 [Bochkovskiy20] on the baseline dataset of 3000 images without our augmentations until convergence.

To test our novel augmentations, we train the same YOLO X model with Mix-up and MOSAIC augmentations on a dataset augmented with the additional *circle grid*, *curtain* and *object relative random box* augmentations proposed here.

The augmented dataset contains roughly an order of magnitude more images which could accelerate training for the same number of epochs. We compensate for this by running both networks to convergence (the point at which further increases in validation set average precision stop).

Evaluation

We evaluate the results with average precision at IOU=0.5. Roughly, we count a prediction successful if the predicted bounding box overlaps a ground truth bounding box by 50% or more.

Result

Our experiments showed that our augmentations significantly improve transfer of a model trained on independent objects with a fixed artificial background to detection in real world scenarios with occlusions. In the simpler case of transferring from fixed artificial backgrounds to realistic scenes average precision at 0.5 IOU increased almost 4 absolute percentage points over state-of-the-art augmentation. In the more challenging case where human hands occlude objects, our augmentations improve average precision by 44 absolute percentage points over state-of-the-art augmentation. Table 2 summarizes the results.

Test Case	Baseline SOA	Our augmentations
Real world images of objects on wooden table	50.6%	**54.5% (+3.9%)**
Real world images + Hand Occlusions	38.7%	**82.7% (+44%)**

Table 1 Experimental results (average precision at IOU=0.50)

To better understand the result, we conducted a qualitative study by inspecting a number of examples. In Figure 8, when we apply the model trained with our augmentation method to scenes with natural backgrounds we see higher confidence in the bounding box a) and c) a lower number of false positives (e.g., the yellow bounding box classifying the table as a cutting board in b) and d).

In Figure 9, when we apply the model trained with our augmentation method to natural scenes with hand occlusions, we see higher confidence on the correct bounding boxes a) and c) and a lower number of false positives compared to the model trained on unaugmented data b) which classified the upper paper towel as a tortilla and d) which failed to detect any towels.

Conclusion

We show that compared to state-of-the-art augmentation techniques like MOSAIC, our foreground and background augmentations can dramatically improve the transfer from models trained on a small set of unoccluded object images taken against a fixed artificial background to real-world recognition of objects in unseen environments with occlusions. Foreground augmentations were particularly important for hand manipulation of objects. The experiments reveal that performance is sensitive to specific types of augmentations and parameters such as coverage. These techniques represent a promising strategy for building robust detectors for specialized industrial contexts where little existing data exists and where obtaining images and expert labels is costly and difficult.

Acknowledgement

This research was funded by Xerox Inc. research funding at the Palo Alto Research Center in Palo Alto. Application testing on kitchen domain funded by DARPA research contract *HR001122C0009.*

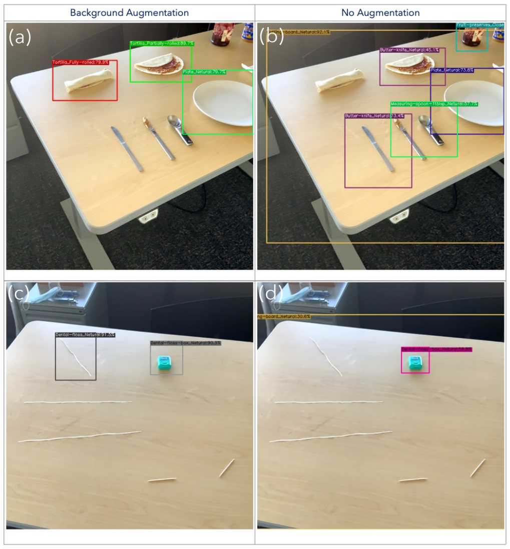


Figure 8 a) the model with augmentation detects tortillas and plate with high confidence b) in same image the model without augmentation misses a tortilla and detects table as a cutting board.

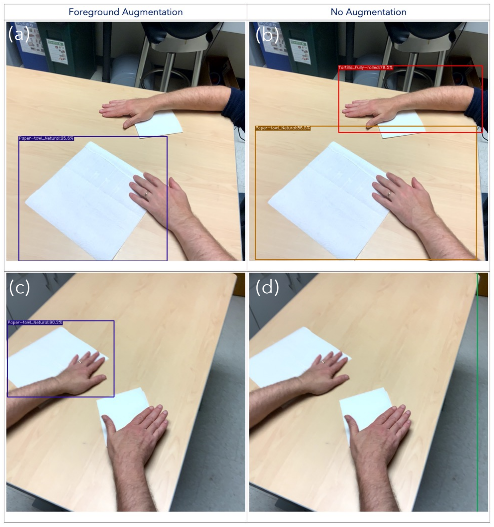


Figure 9 a) model trained with augmentation detects the paper towel with high confidence. b) in same image the model without augmentation detects lower paper towel with lower confidence and second paper towel is mistaken for tortilla. c) model with augmentation detects occluded paper towel d) in same image the model without augmentation fails to detect any towels.

Author Biography

Yan-Ming Chiou *is a member of scientific staff at the Palo Alto Research Center, where he explores augmented reality and computer vision applied to AR assistance systems. He received his Ph.D. from the University of Delaware.*

Bob Price *is a Principal Scientist at the Palo Alto Research Center where he investigates probabilistic and deep learning models with custom structures to recover spatial and relational structure of scenes and activities. He got his PhD from the University of British Columbia and did a Post Doctoral Fellowship at the University of Alberta.*

References

[1] [Bochkovskiy20] Alexey Bochkovskiy*, Chien-Yao Wang*, Hong-Yuan Mark Liao, (2020) YOLOv4: Optimal Speed and Accuracy of Object Detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Online virtual hosting.pp.126-138.

[2] [DeVires17] Terrance DeVries, Graham W. Taylor. Improved Regularization of Convolutional Neural Networks with Cutout. 2017. https://arxiv.org/abs/1708.04552

[3] [Hao20] Wang Hao, Song Zhili. Improved Mosaic: Algorithms for more Complex Images. Journal of Physics: Conference Series AINIT 2020.

[4] [Sangdoo19] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo, (2019). CutMix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) . Seoul, Korea. pp. 6023–6032. WARHOL: Wearable holographic object labeler

[5] [Shreve20] Matthew Shreve, Bob Price, Les Nelson, Raja Bala, Jin Sun, Srichiran Kumar. WARHOL: Wearable Holographic Object Labeler, Electronic Imaging, Vol 32, pp 1-10, 2020

[6] [Yuan21] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, Pengchuan Zhang. Florence: A New Foundation Model for Computer Vision.

[7] [Zhang17] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, David Lopez-Paz. Mixup: Beyond Empirical Risk Minimization. ICLR 2017

[8] [Zheng21] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, Jian Sun. YOLOX: Exceeding YOLO Series in 2021 https://arxiv.org/abs/2107.08430