Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection

Carnegie Mellon University

Debidatta Dwibedi Ishan Misra Martial Hebert

Carnegie Mellon University

THE ROBOTICS INSTITUTE

Overview

Goal

- Create annotated datasets for instance detection with minimal manual effort

Key Ideas

- For feature rich object instances, local context matters more than global context
- Cut real instance masks and paste on real backgrounds to create data
- Multiple blending modes reduce pasting artifacts

Outcome

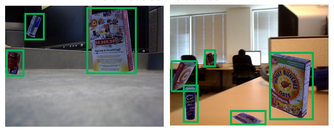
- Simple and fast method to generate images. Outperforms slower complex methods.
- 10% real + synthetic data performs as well as 100% real data.

Cut

Object Instances Background Scenes

Paste

Generated Scenes



Learn

Detections on Real Images

Instance Detection

Object Detection

Granola Bars		Cups			
					
Granola Bar 1	Granola Bar 2	Cup 1	Cup 2	Cup 3	Cup 4

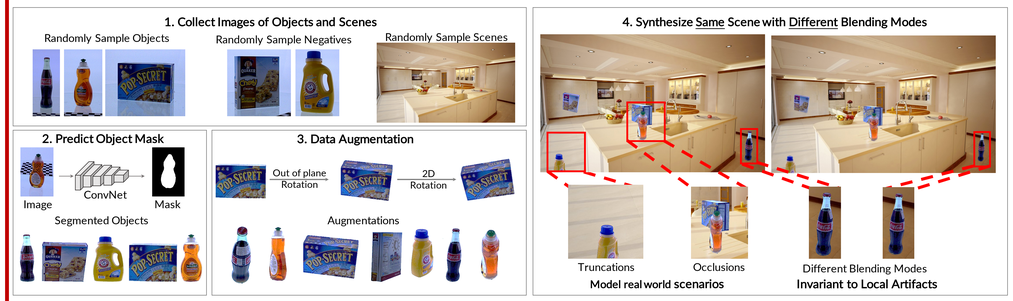
Instance Detection

Problems with manual data collection

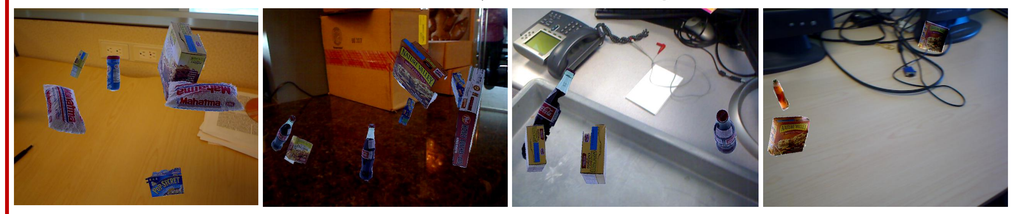
- Physically creating and capturing scenes for instance detection is a major bottleneck
- Ensuring visual diversity in terms of views, occlusions, backgrounds and lighting is challenging
- Generalization to new environments is hard with limited data

Easy misses by object detector trained on real data

Approach



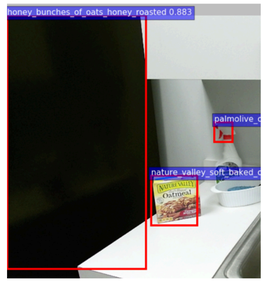
Examples of Synthesized Images



Blending Modes

- Naïve pasting of objects results in pixel artifacts
- Synthesizing same scene with multiple blending modes improves performance

No Blending Gaussian Blurring Poisson Blending



Typical false positives seen when trained with no blending

Patch-level Realism

- Difficult to model global realism
- For feature rich regions, local context matters more than global context.
- Patch-level realism is easy and produces useful training data

Patches that look realistic in unrealistic scenes

Which synthesizing factors matter the most?

Data Augmentation	w/o Distractor Objects	w/o Blending	w/o 2D Rotation	w/o 3D Rotation	w/o Truncation	w/o Occlusion	w Distractor Objects
mAP	73.7	-7.8	-4.0	-5.4	-1.9	-10.6	+2.5

Blending and Occlusion modeling give the most performance boost

Experiments

Object Instances	BigBIRD
Random Backgrounds	UW RGBD Dataset
Real Images	GMU Kitchen Scenes
Unseen Scenes Dataset	Active Vision Dataset
Object Detection Model	Faster R-CNN

Comparison to existing approaches

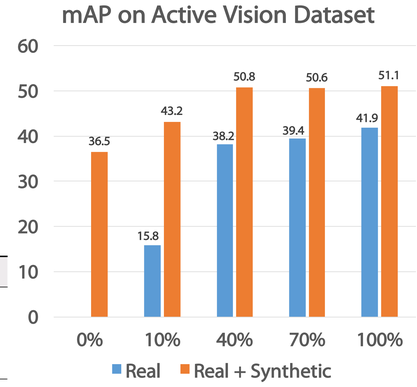
- Our approach outperforms existing approaches that consider geometry and semantics while synthesizing scenes
- Local features matter more because the model considers features in a region

Dataset	Real Images from GMU	Semantic-and-Geometry Aware Synthesis	Synthetic Images (Ours)	Semantic-and-Geometry Aware Synthesis + Real	**Synthetic Images (Ours) + Real Images**
mAP	86.3	51.7	76.2	85.0	**88.8**

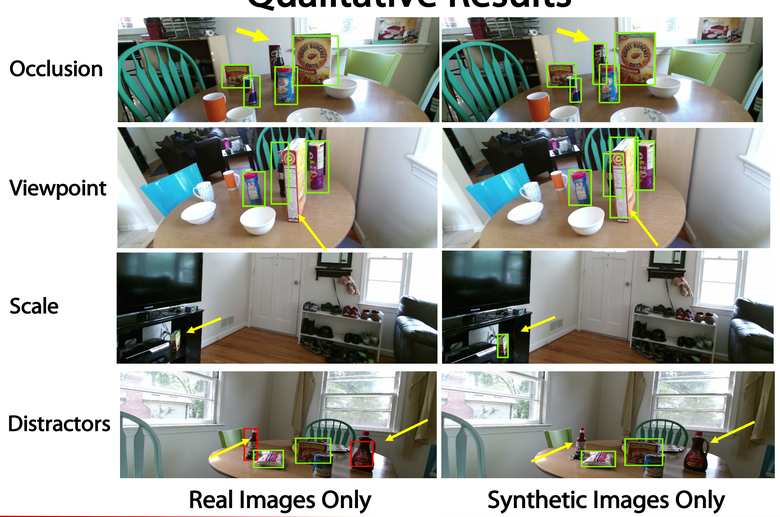
Synthetic data vs. Real Data

- There might be a degradation of performance of models trained with real images on new datasets
- Synthetic data contains complementary information to real images and doesn't hurt performance

Dataset	mAP
Real Images from GMU	41.9
Synthetic Images	36.5
Synthetic Images + Real Images	**51.1**



Qualitative Results



Code available at: https://goo.gl/imXRt7