SuctionNet: An end to end model for suction

Nikita Demir
ndemir@stanford.edu

Jose Miguel Giron
jmgiron@stanford.edu

Mauricio Wulfovich
mauriw@stanford.edu

Motivation

- In order for robots to robustly perform useful tasks in the real-world--from stocking grocery shelves to assembling complex machinery--they must be able to interact with varied previously-unseen objects
- A crucial and necessary first component of such tasks is for robots to automatically detect the best places to grab these items. In this paper, we worked with a synthetic dataset provided by Nimble.ai created to train robots that use suction grasping.

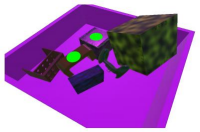
Problem Definition

- Input: RGB-D images of objects labeled with their optimal suction location
- Output: predicted optimal suction grasp labels on unseen RGB-D images
- Goal: End-to-end system

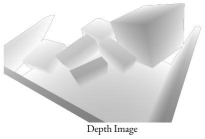
Related Works

- Long et al. "Fully Convolutional Networks for Semantic Segmentation" 2014.
- Pan et al."Shallow and Deep Convolutional Networks for Saliency Prediction" 2016.
- James and Wohlhar et al "Sim-to-Real via Sim-to-Sim: Data-efficient Robotic Grasping via Randomized-to-Canonical Adaptation Networks" 2018.
- Wang et al. "Depth-aware CNN for RGB-D Segmentation" 2018.
- Eitel et al. "Multimodal Deep Learning for Robust RGB-D Object Recognition" 2015.

Data



Correct output (in green) overlayed on RGB Image

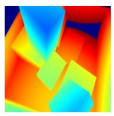


Depth Image

- Synthetic dataset provided to us by Nimble.ai generated by dropping randomly selected 3D objects on a simulated tray using a physics simulator.

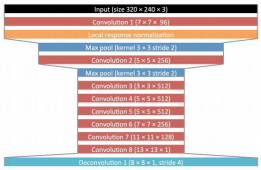
Processing

- Resized to 224x224, strip alpha dimension, divide by 255
- Label: every pixel either 0 (not good suction) or 1 (good suction)
- Depth: Jet-encoding maps near pixels to red, over green, all the way to further pixels which map to green



Processed depth image

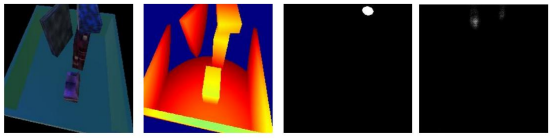
Model



RGB FCN

RGB Encoder	Depth Encoder	
input (224 x 224 RGB image)	input (224 x 224 Jet image)	
conv3-64 conv3-64	conv3-64 conv3-64	Layer 1
maxpool = layer 1 output		
conv3-128 conv3-128	conv3-128 conv3-128	Layer 2
maxpool = layer 2 output		
conv3-256 conv3-256 conv3-256	conv3-256 conv3-256 conv3-256	Layer 3
maxpool = layer 3 output		
conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512	Layer 4
maxpool = layer 4 output		
conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512	Layer 5
maxpool = layer 5 output		
conv1-4096	conv1-4096	Layer 6
conv1-4096	conv1-4096	Layer 7
Decoder		
input (avg layer 7 output)		
conv1-3		Layer 8
deconv4-512		Layer 9
Add avg layer 4 output		Skip connection
deconv4-256		Layer 10
Add avg layer 3 output		Skip connection
deconv16-4		Layer 11

Results and Analysis



Input RGB | Input Depth | Ground Truth | Our Prediction

	Shallow SalNet	Deep SalNet	RGB FCN-8s	Depth FCN-8s	RGB-D FCN-8s
Intersection Over Union	0.00	0.00	0.15	0.12	0.13

Input RGB | Ground Truth | Our Prediction

Input RGB | Ground Truth | Our Prediction

- Our best model, RGB FCN-8s, manages to capture general good suction points but not necessarily the best ones as captured in the ground truth label

Future Work

- Fine-tuning on real data
- Experimenting with different merge strategies for the FCNN
- Manual learning rate decay