Deep Object Pose Estimation for semantic Robotic Grasping of Household Objects – CoRL 2018

Jonathan Tremblay Yu Xiang, Thang To Dieter Fox, Balakumar Sundaralingam Stan Birchfield - NVIDIA

KIST
송명하

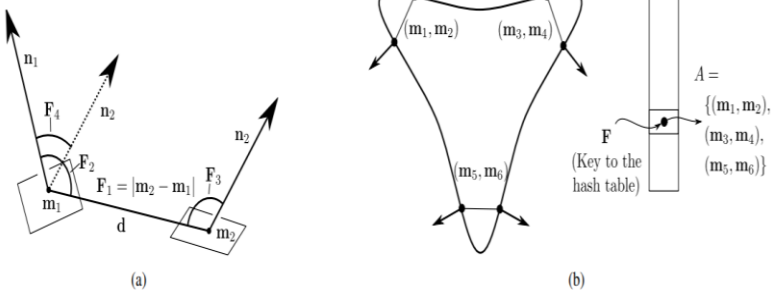
Content

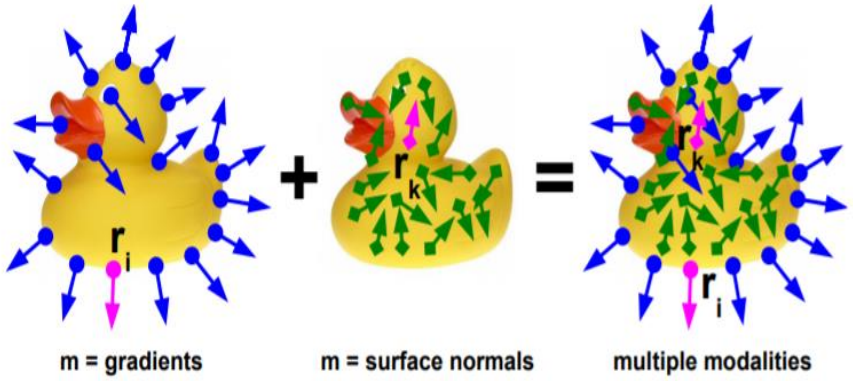
1. Introduction

1. Introduction

1. Introduction

1. Introduction





Model Globally, Match Locally:
Efficient and Robust 3D Object Recognition (CVPR2010)

Multimodal Templates for Real-Time Detection of Texture-less Objects in Heavily Cluttered Scenes(ICCV2011)

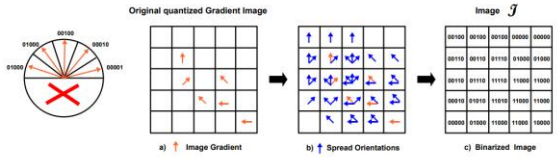
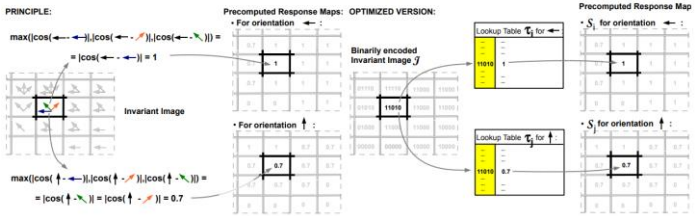
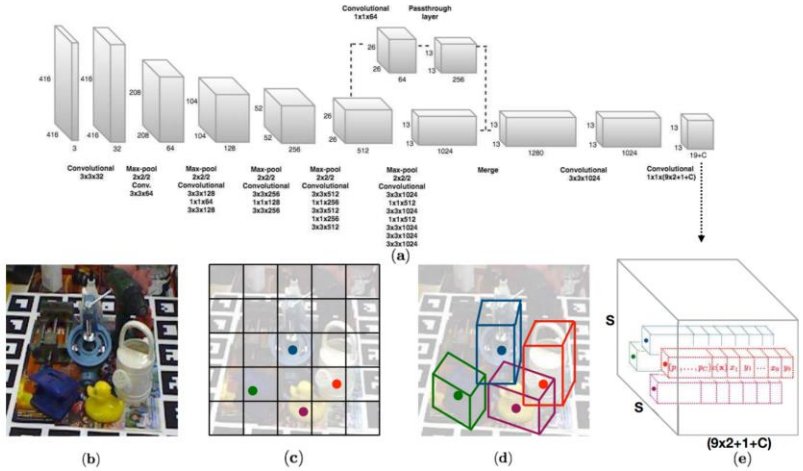


Fig. 4: Spreading the gradient orientations. **Left:** The gradient orientations and their binary code. We do not consider the direction of the gradients. **a)** The gradient orientations in the input image, shown in orange, are first extracted and quantized. **b)** Then, the locations around each orientation are also labeled with this orientation, as shown by the blue arrows. This allows our similarity measure to be robust to small translations and deformations. **c)** $\mathcal{J}$ is an efficient representation of the orientations after this operation, and can be computed very quickly. For this figure, $T = 3$ and $n_o = 5$. In practice, we use $T = 8$ and $n_o = 8$.

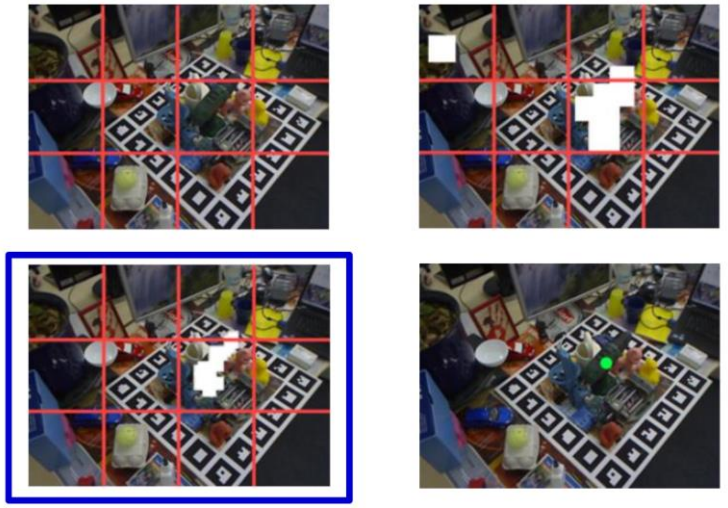


Gradient Response Maps for Real-Time Detection of Texture-Less Objects (PAMI 2012)

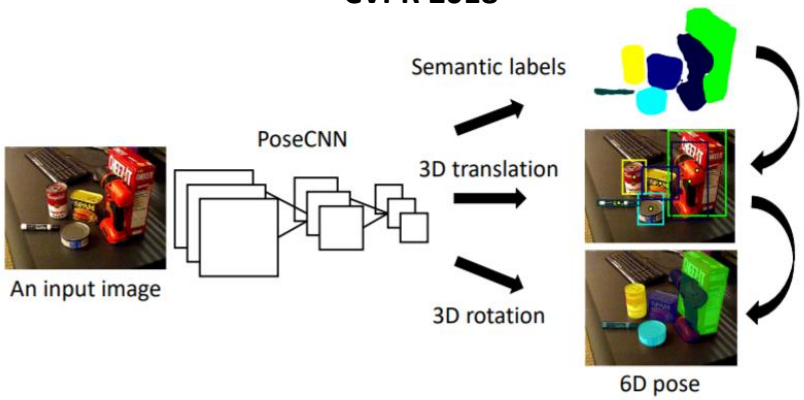
1. Introduction



Real-Time Seamless Single Shot 6D Object Pose Prediction -CVPR 2018

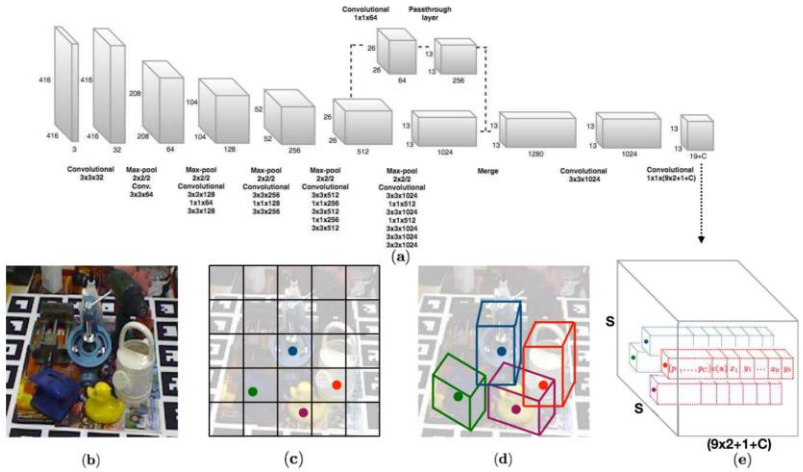


bb8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without Using Depth - 2017 ICCV

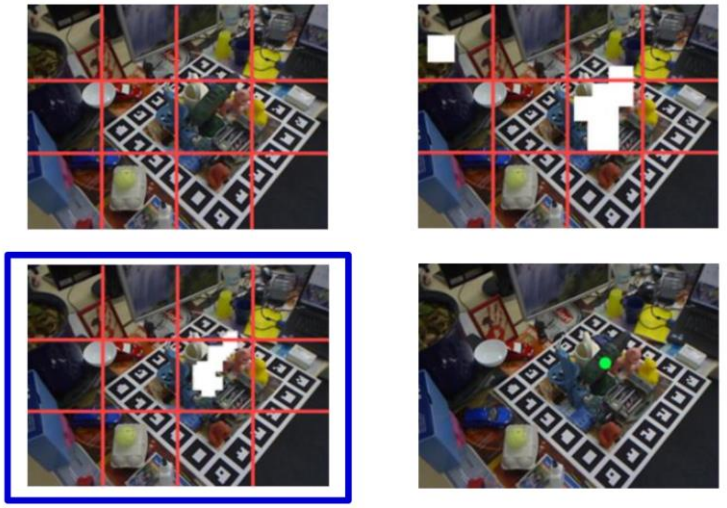


PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes – RSS 2018

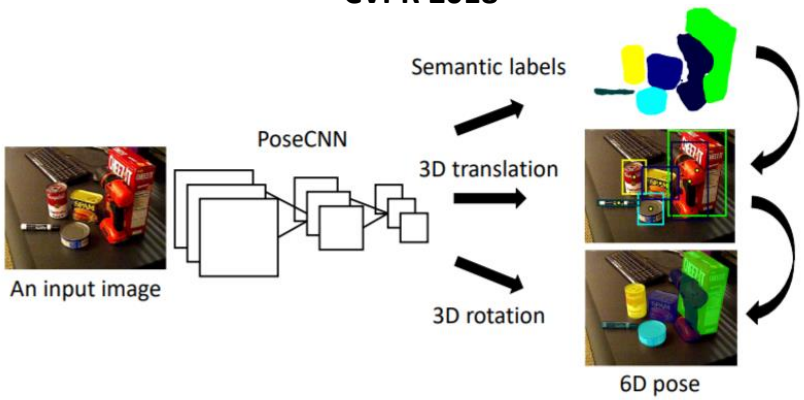
1. Introduction



Real-Time Seamless Single Shot 6D Object Pose Prediction -CVPR 2018



bb8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without Using Depth - 2017 ICCV

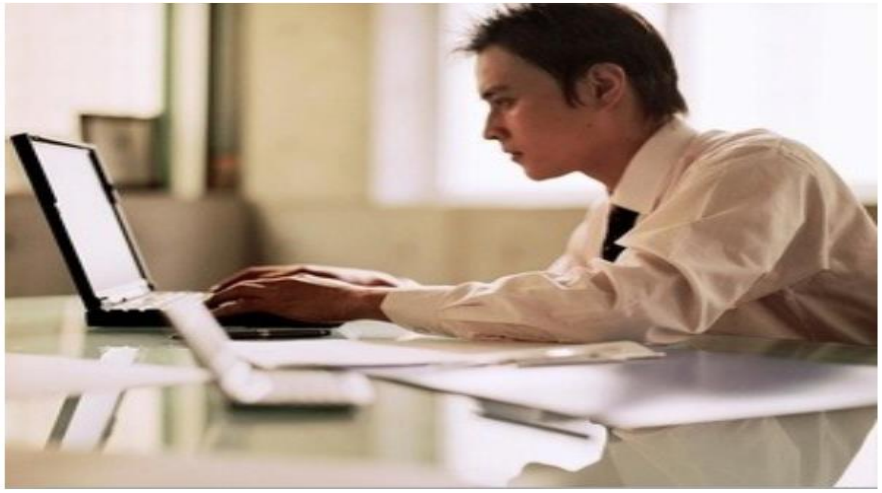


PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes – RSS 2018

데이터 확보가 문제

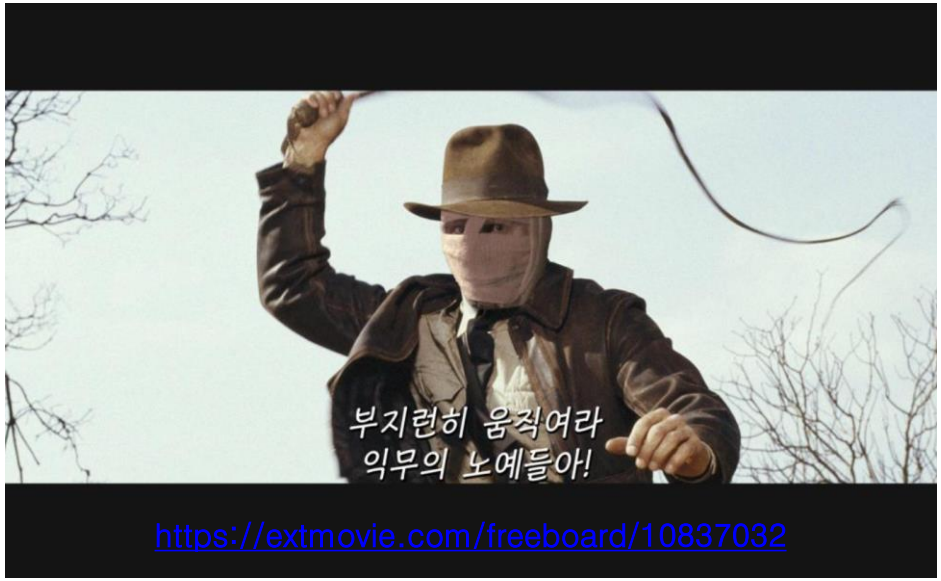
1. Introduction

https://extmovie.com/freeboard/10837032

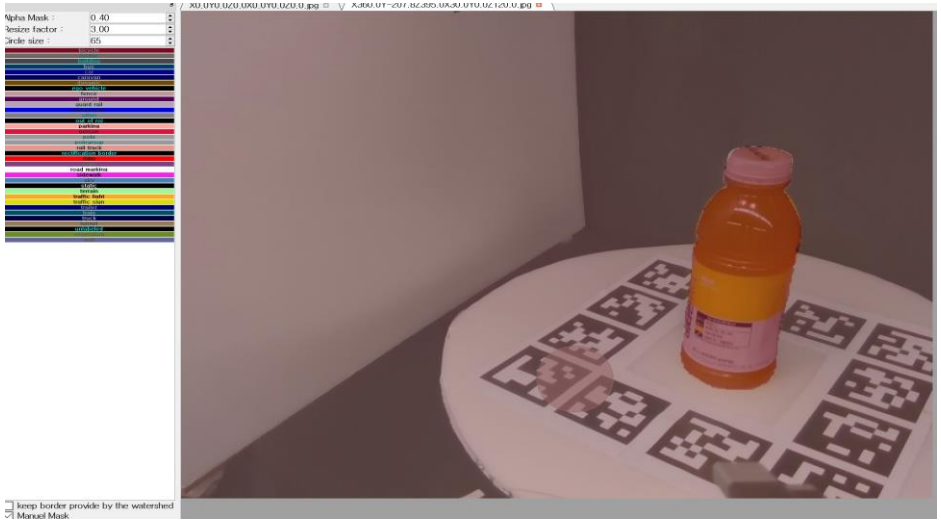


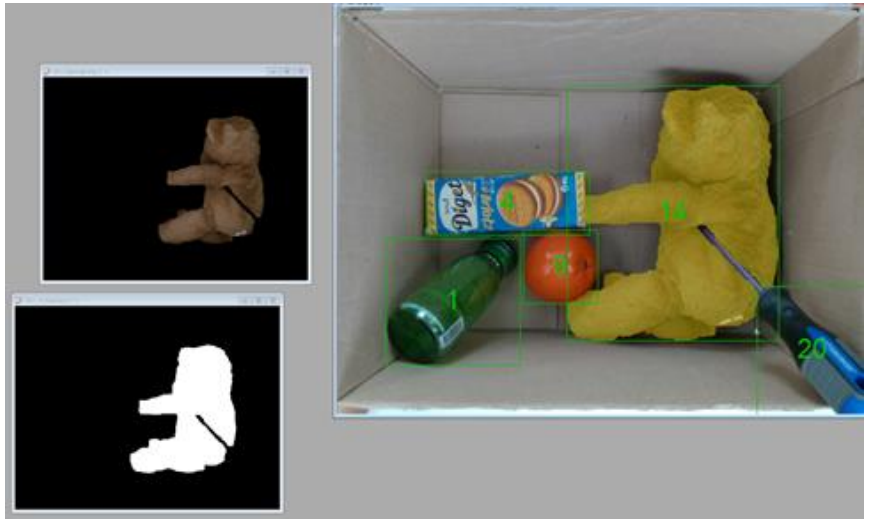
http://blog.daum.net/_blog/BlogTypeView.do?blogid=0UGOf&articleno=132&categoryId=24&regdt=20111019122204

1. Introduction



https://extmovie.com/freeboard/10837032





1. Introduction

Synthetic Data를 생성하여 data부족 문제를 해결하고 있음.

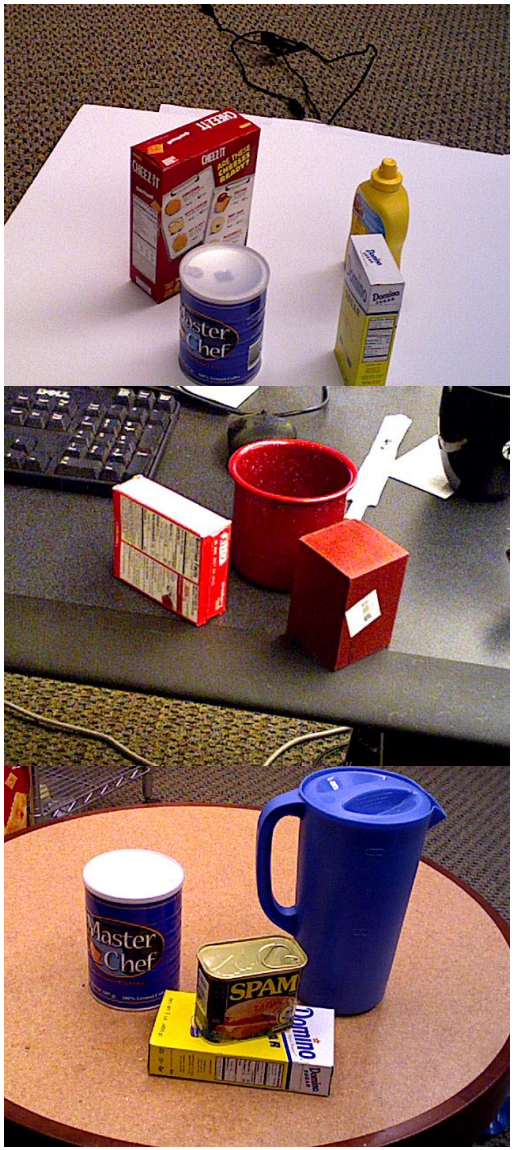
1. Introduction

1. Introduction

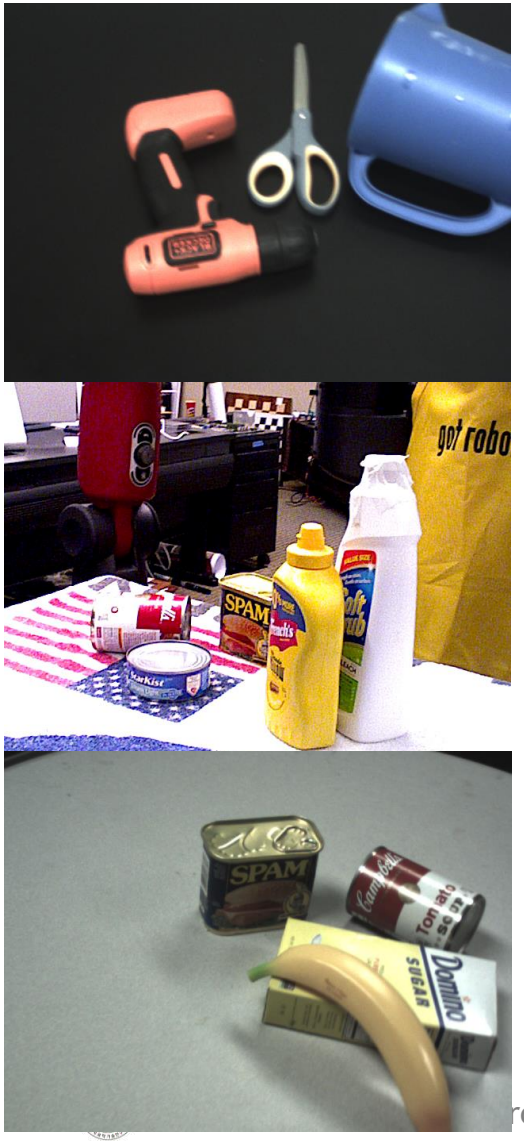
≠

Reality Gap

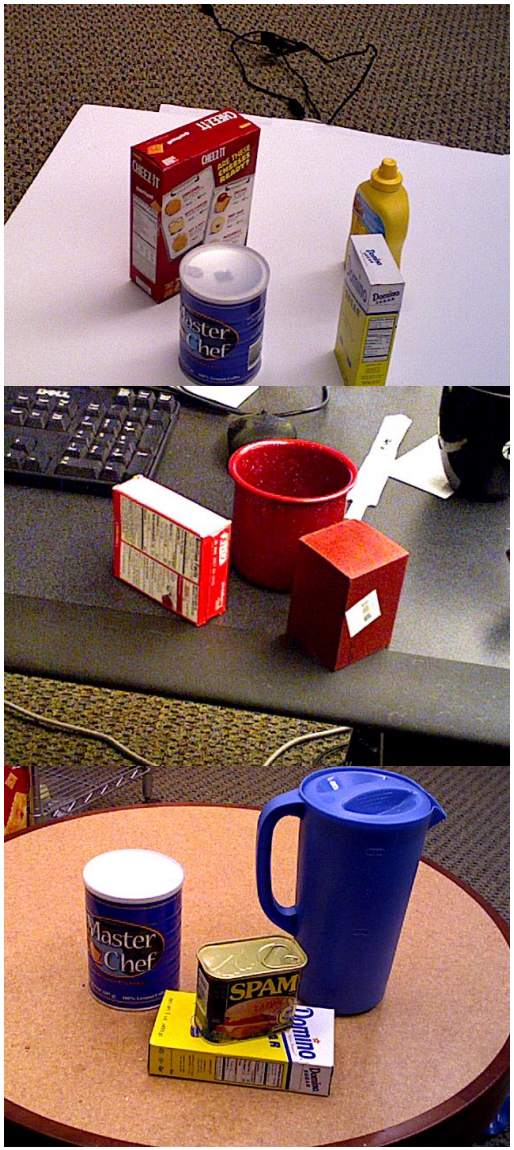
1. Introduction



Fixed Scene



1. Introduction



≠

Fixed Scene

Reality Gap

1. Introduction

Contribution

- A one-shot, deep neural network-based system that infers, in near real time, the 3D poses of known objects in clutter from a single RGB image without requiring post-alignment. This system uses a simple deep network architecture, trained entirely on simulated data, to infer the 2D image coordinates of projected 3D bounding boxes, followed by perspective-n-point (PnP) [12]. We call our system DOPE (for "deep object pose estimation").

- Demonstration that combining both non-photorealistic (domain randomized) and photorealistic synthetic data for training robust deep neural networks successfully bridges the reality gap for real-world applications, achieving performance comparable with state-of-the-art networks trained on real data.

- An integrated robotic system that shows the estimated poses are of sufficient accuracy to solve real-world tasks such as pick-and-place, object handoff, and path following.

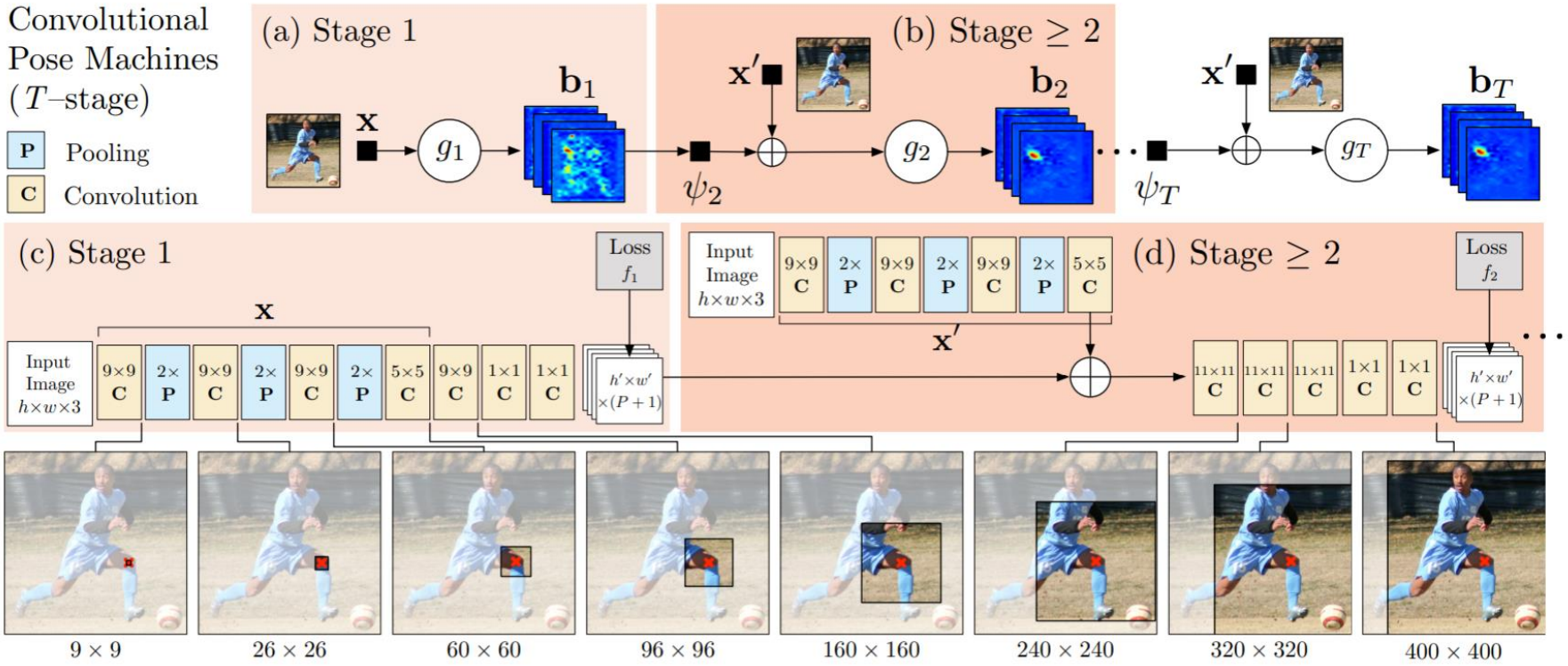
2. Approach

- **Network**

- **Data generation**

2. Approach – Network

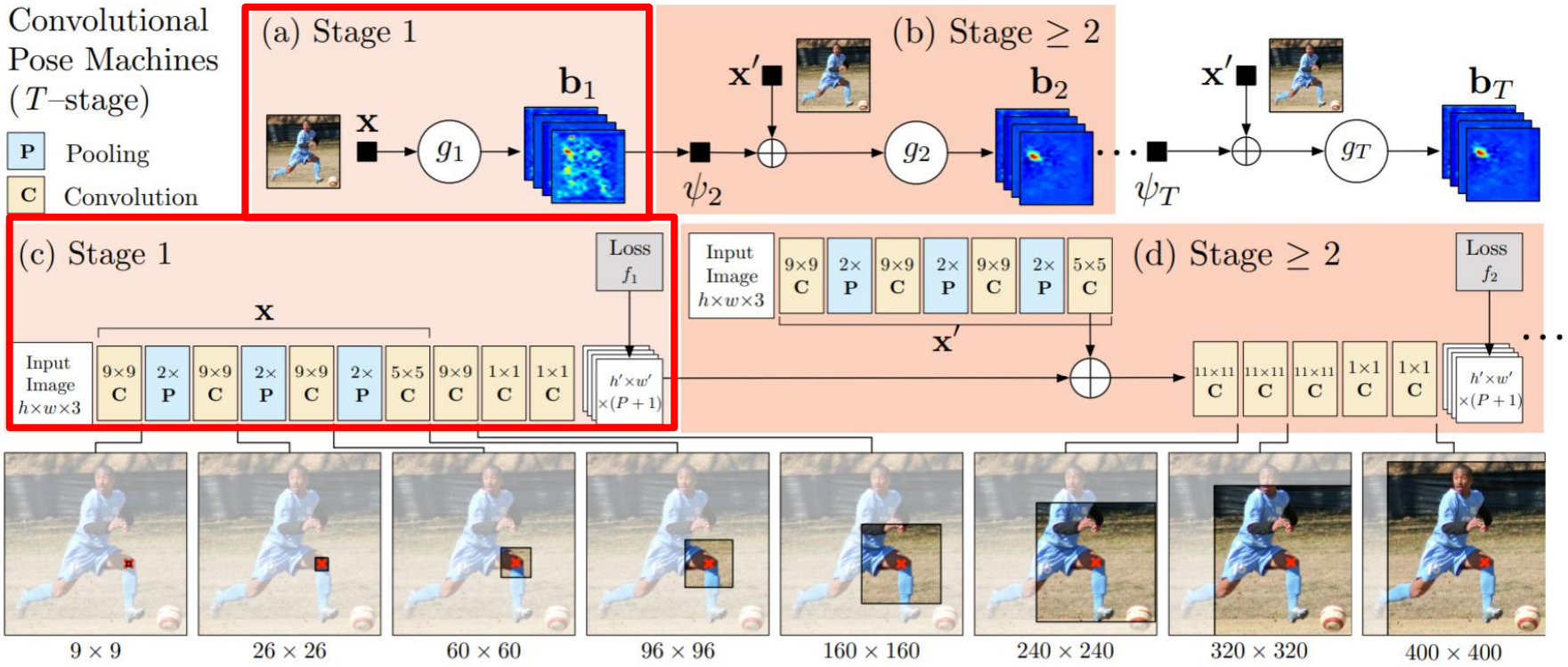
Input : Single RGB image



Convolutional Pose Machines – CVPR 2016

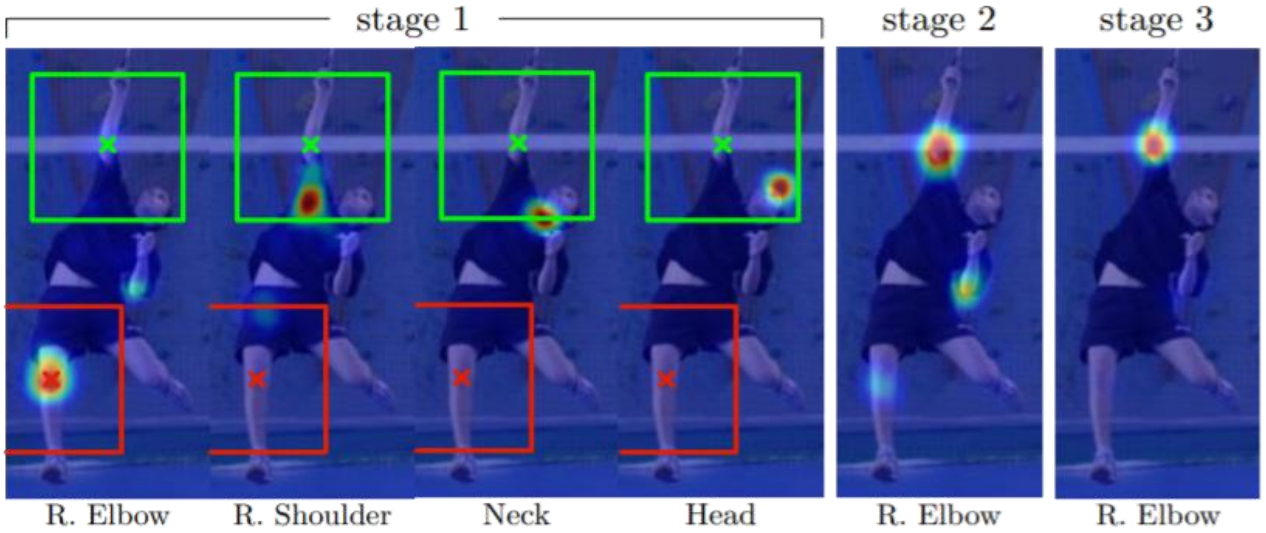
2. Approach – Network

Input : Single RGB image



Convolutional Pose Machines – CVPR 2016

2. Approach - Network



Convolutional Pose Machines – CVPR 2016

Network 설계 이유 :

초반 stage에서는 Local한 이미지 영역을 통해 관절 찾음.

후반 stage에서는 더 커진 Receptive Field의 영향으로 서로 다른 관절 간의 관계까지 고려되므로 더욱 정확한 heatmap을 얻을 수 있음.

2. Approach - Network

VGG19

stage1

stage2

stag3

stage4

stage5

stage6

2. Approach - Network

VGG19

stage1

stage2

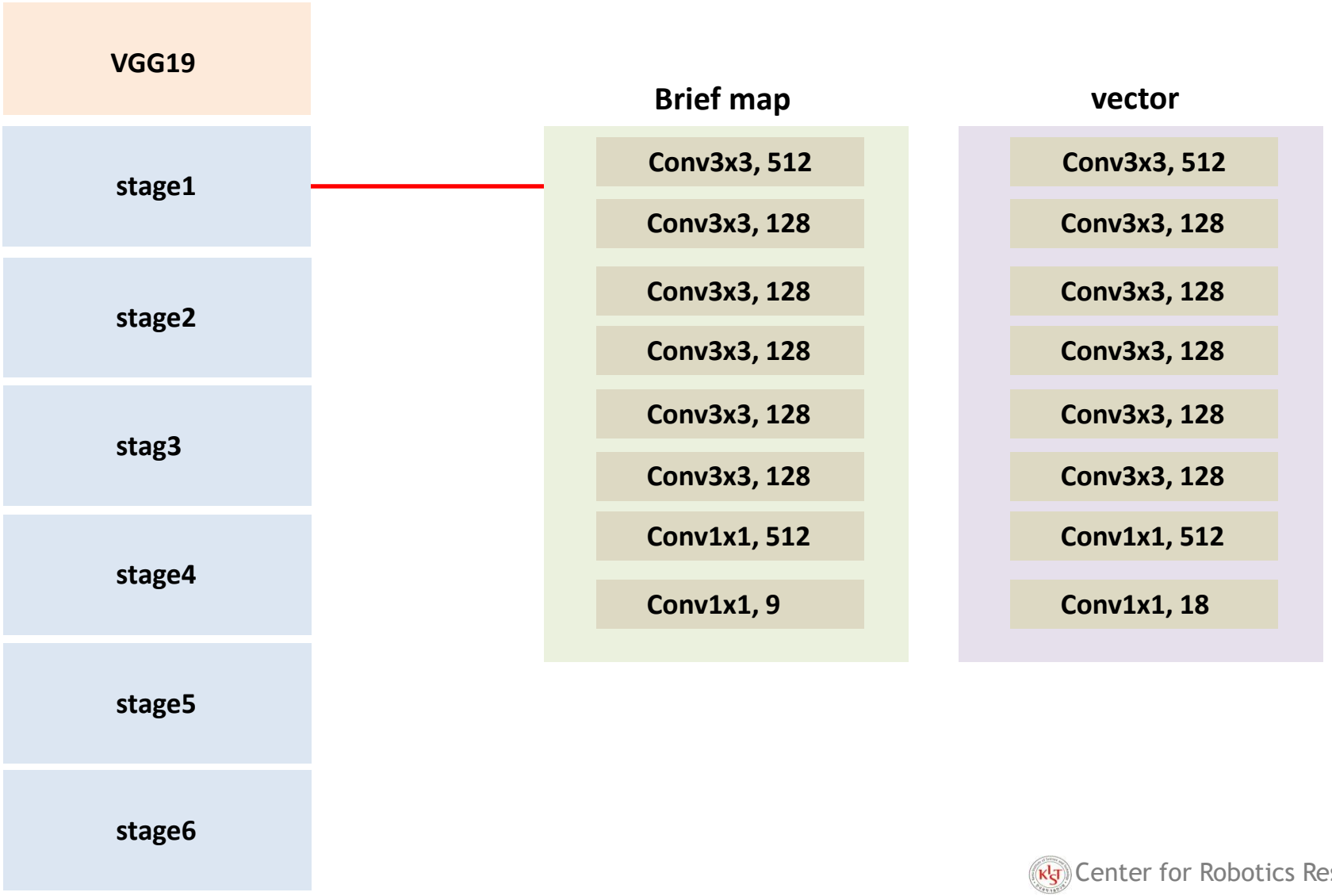
stag3

stage4

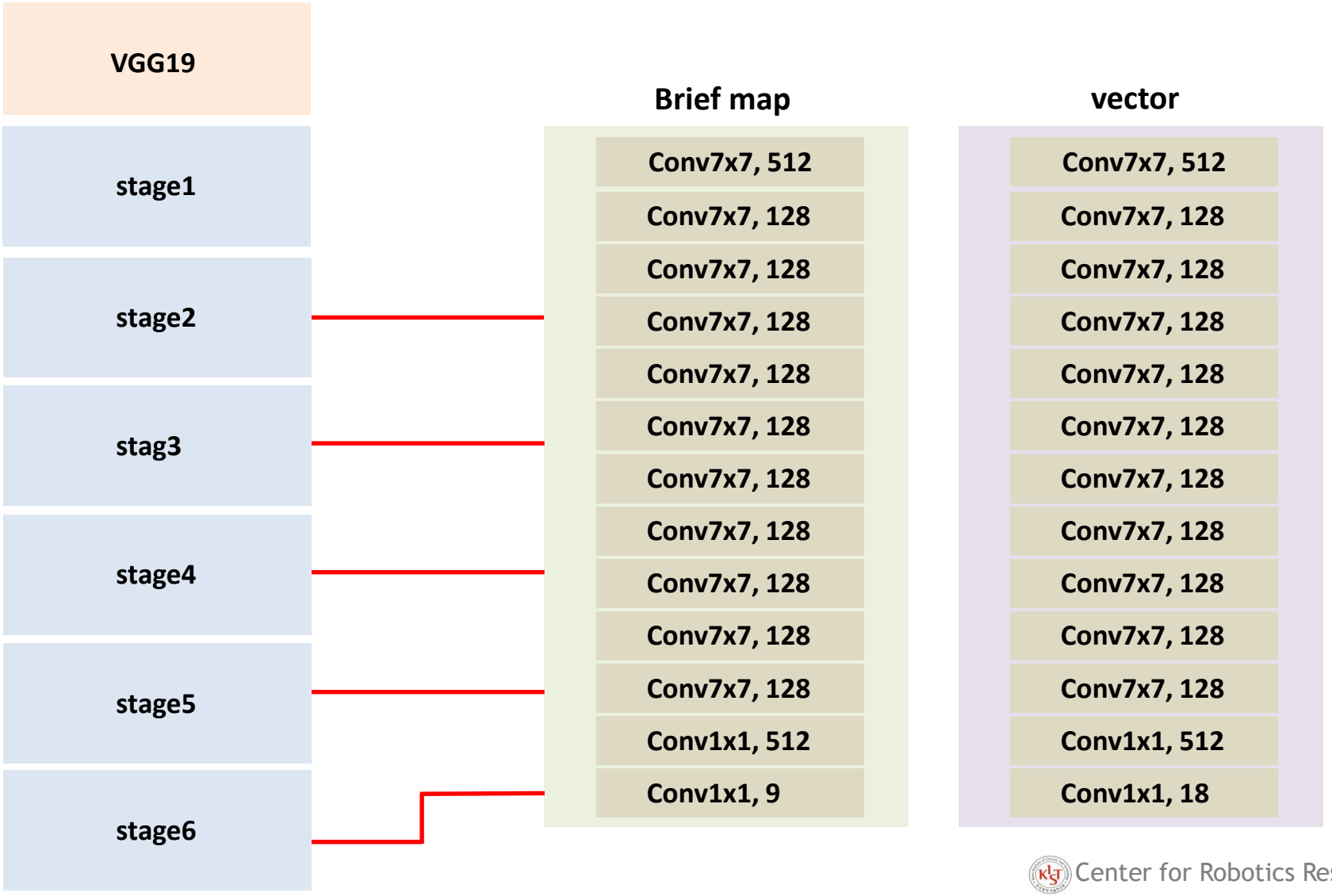
stage5

stage6

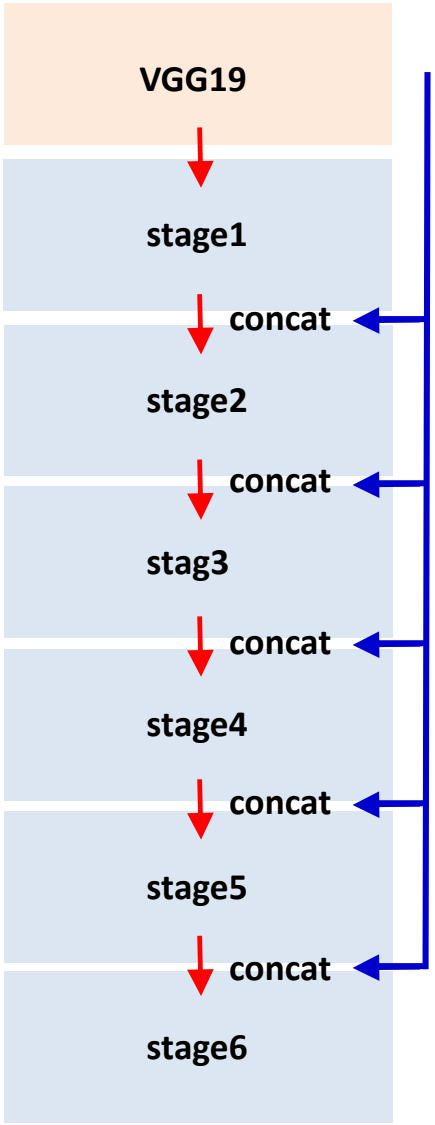
2. Approach - Network



2. Approach - Network



2. Approach - Network



2. Approach - Network

VGG19

stage1

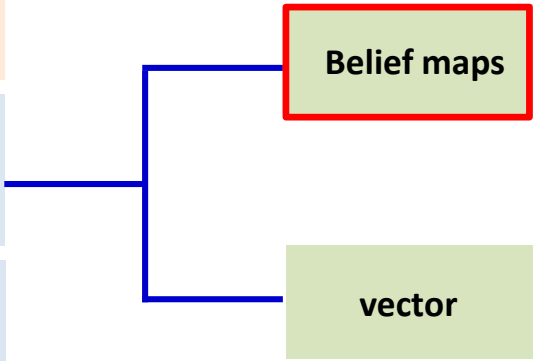
stage2

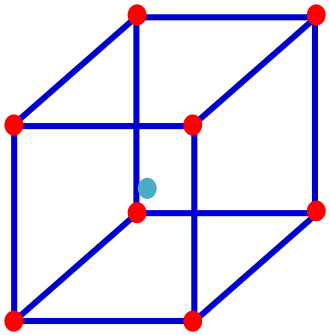
stag3

stage4

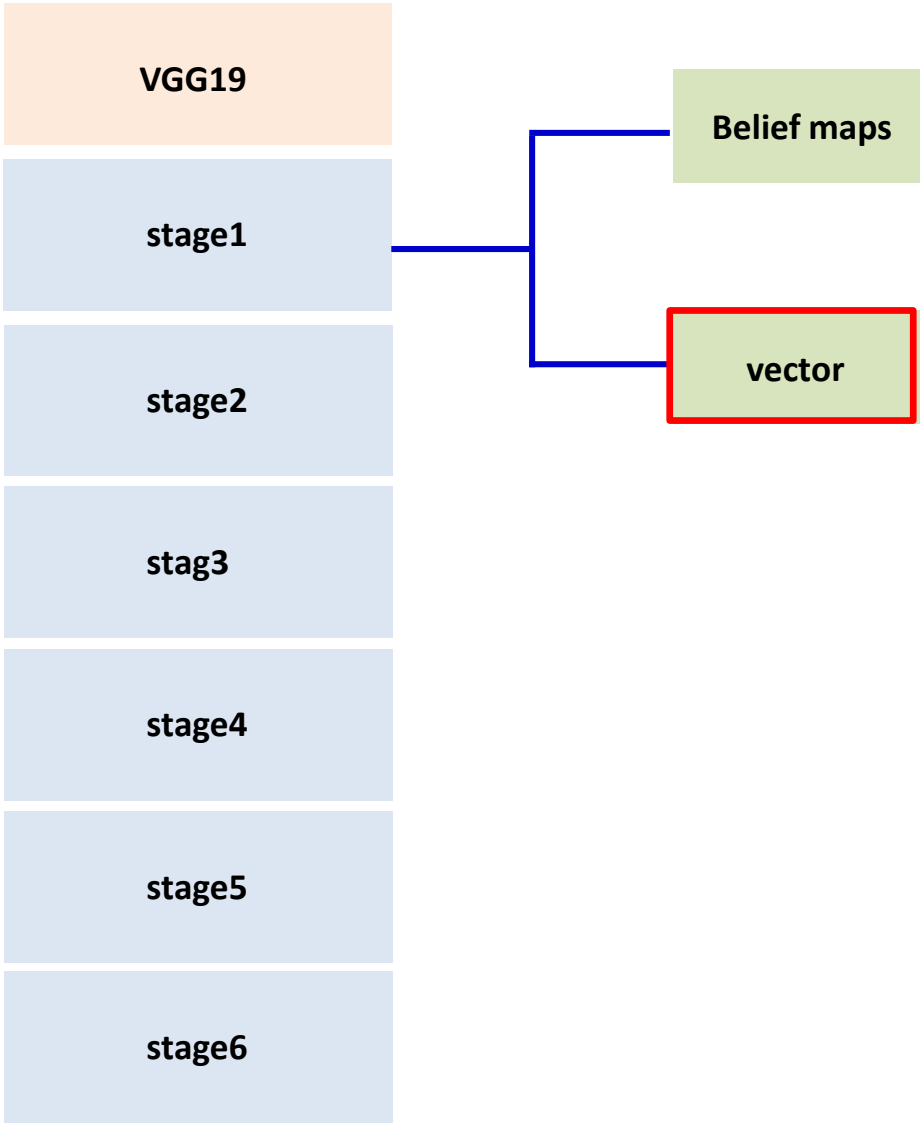
stage5

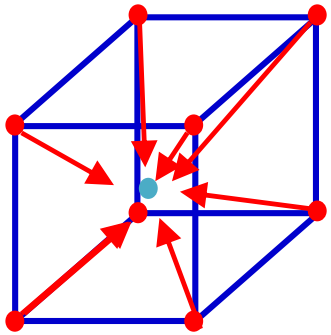
stage6





2. Approach - Network





2. Approach - Network

VGG19

stage1

stage2

stag3

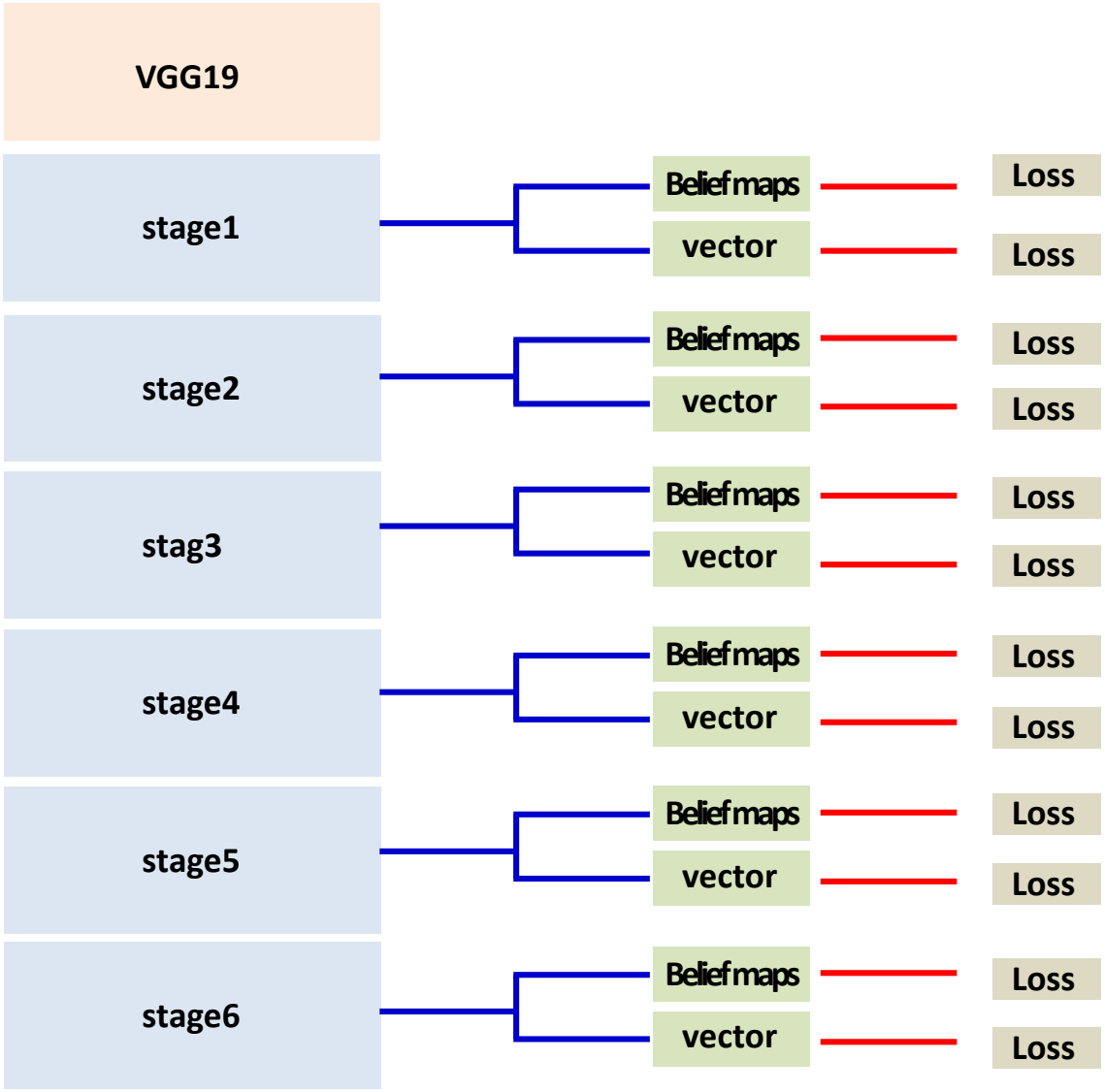
stage4

stage5

stage6

Vanishing Gradient

2. Approach - Network

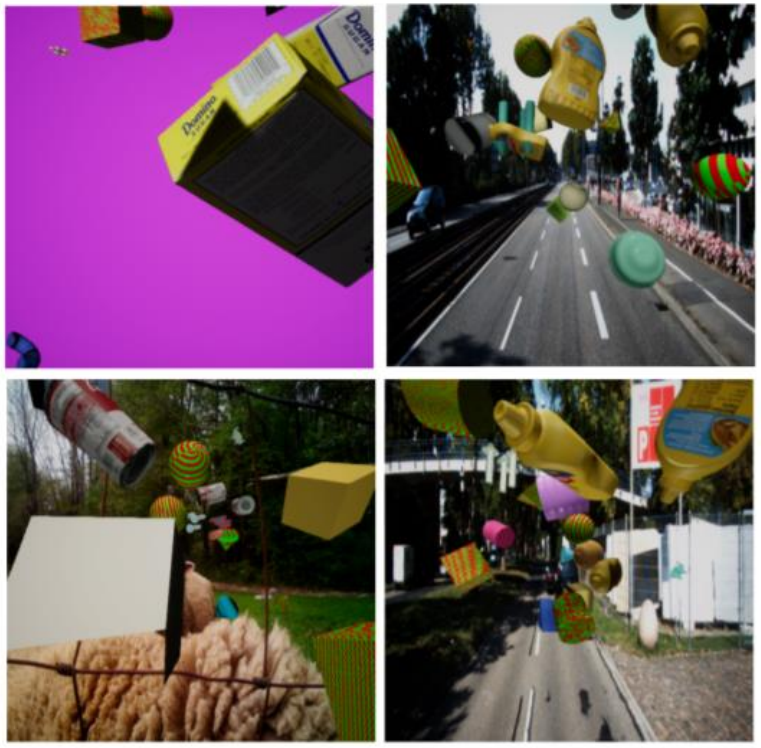


2. Approach – Data Generation

2. Approach – Data Generation

Reality Gap

2. Approach – Data Generation



Non-photorealistic
(Dain Randomization)

photorealistic

2. Approach – Data Generation

Non-photorealistic(Domain Randomization)

3D Model(Fore Ground)

CoCo dataset 2D image
(Back Ground)

2. Approach – Data Generation

Non-Photorealistic

2. Approach – Data Generation

photorealistic

1. Unreal Engine4(UE4) called NDDS
2. Foreground objects in 3D background scenes with physical constraint
3. While the objects were falling, the virtual camera system was rapidly teleported to random azimuths, elevations, and distances with respect to a fixation point to collect data. Azimuth ranged from –120◦ to +120◦ (to avoid collision with the wall, when present), elevation from 5 ◦ to 85◦ , and distance from 0.5 m to 1.5 m

3. Experimental Results

Datasets : YCB Video Dataset

YCB 21개 물체 중 5개(cracker box, sugar box, tomato soup can, mustard bottle, potted meat can)을 사용.

Logitech C960 camera로 Light condition을 다양하게 해서 수집한 dataset

Test : 2949 frames

3. Experimental Results

- **For training**

  ~60k non-photorealistic data + ~ 60k photorealistic image frames.

  Belief maps와 vector에 L2 Loss를 적용.

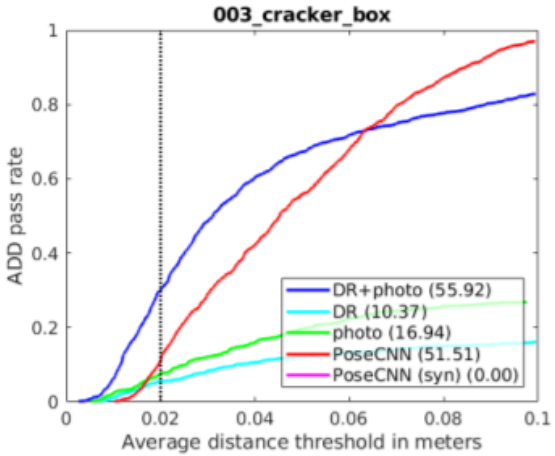
- **Metric**

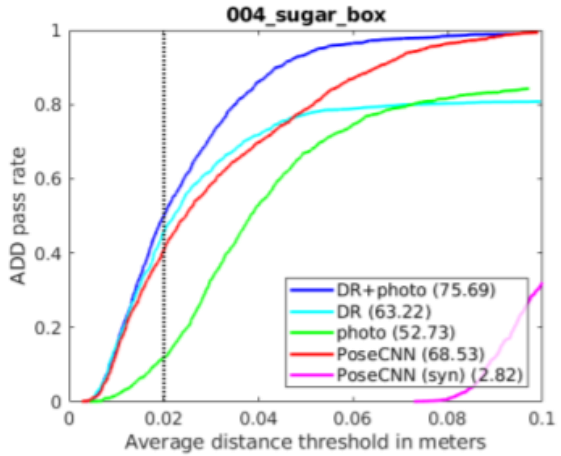
  ADD(average distance) 3D Point 간 Euclidean distance

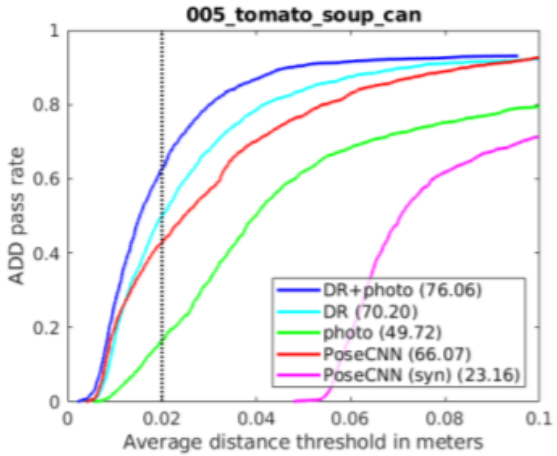
- **비교 모델**

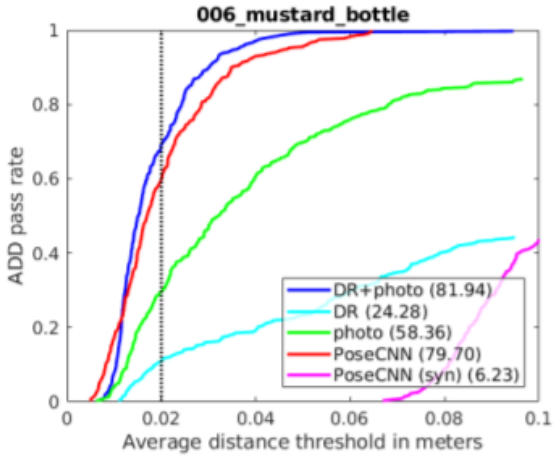
  PoseCNN(SOTA) : real data + synthetic data 로 학습시킨 모델.

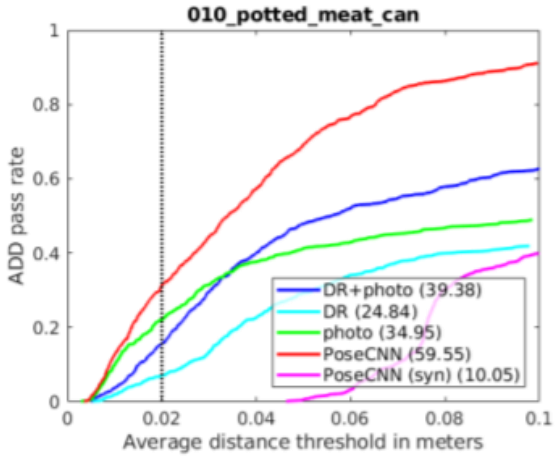
3. Experimental Results



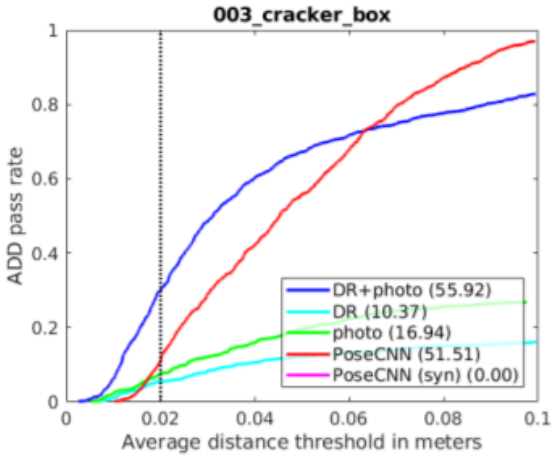


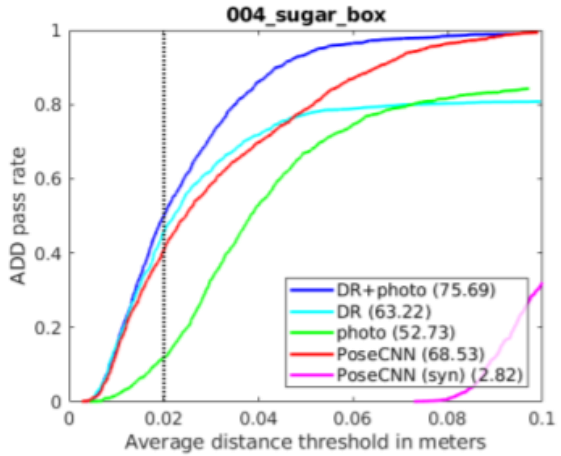


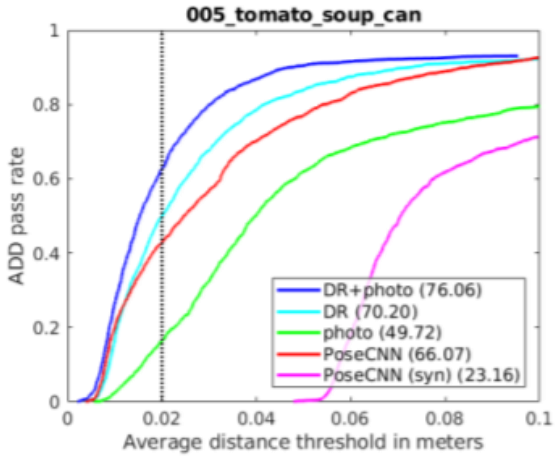


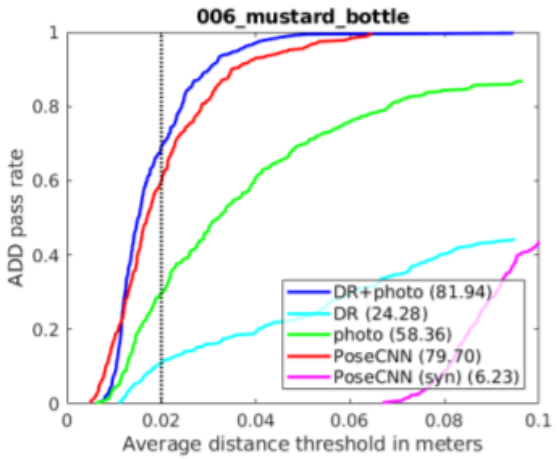


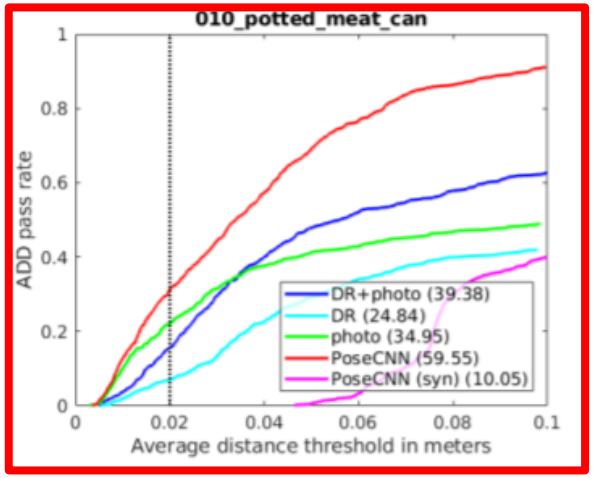
3. Experimental Results



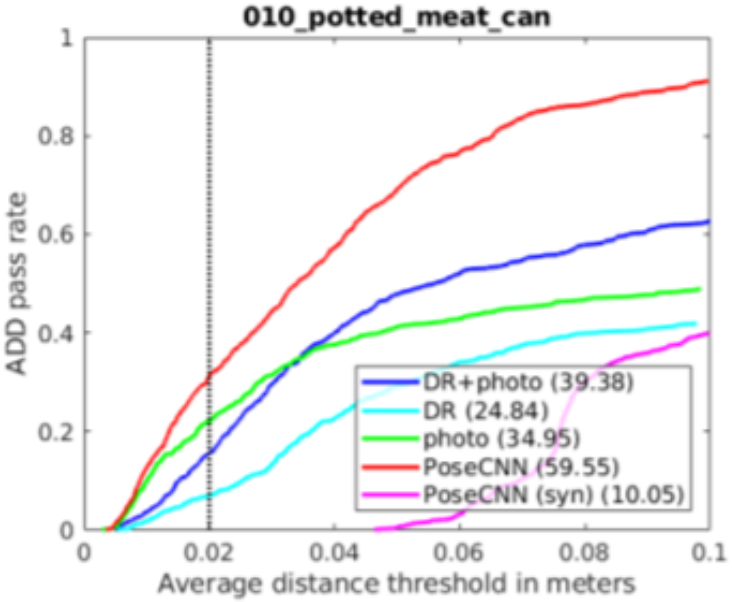








3. Experimental Results

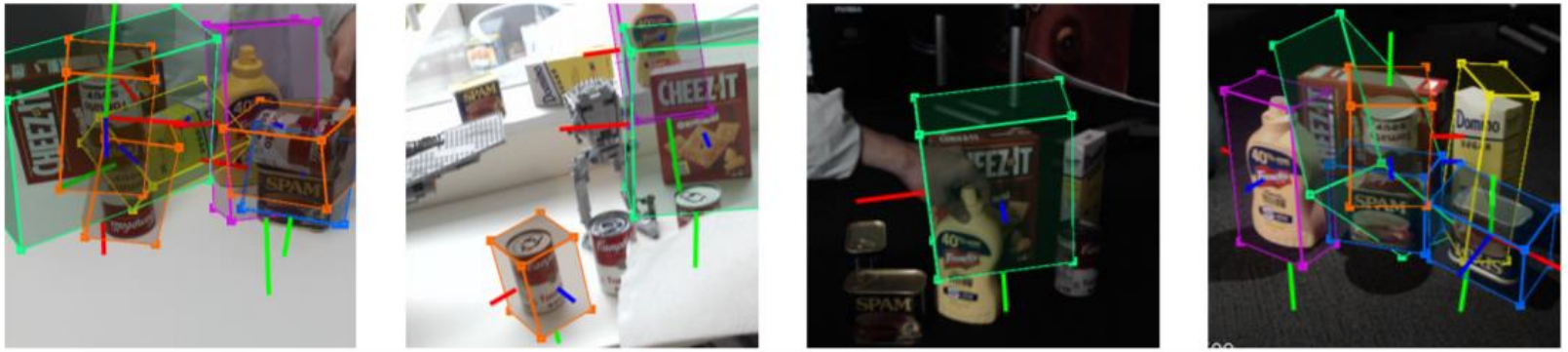


The reason our network fails to detect many of the potted meat can instances is due to severely occluded frames in which only the top of the can is visible; since our synthetic data does not properly model the highly reflective metallic material of the top surface, the network does not recognize these pixels as belonging to the can. We leave the incorporation of such material properties to future work.

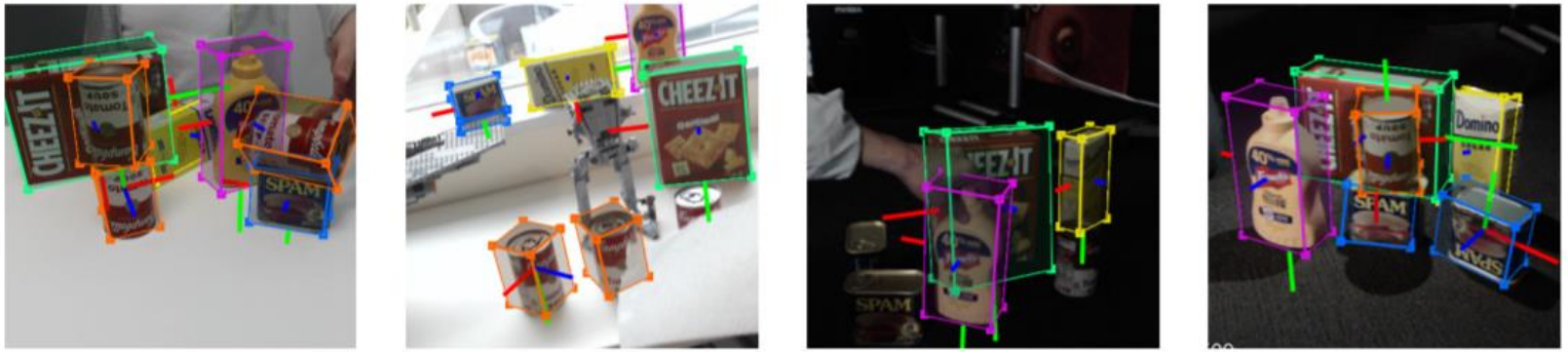
3. Experimental Results

Extreme lighting condition

PoseCNN [5]

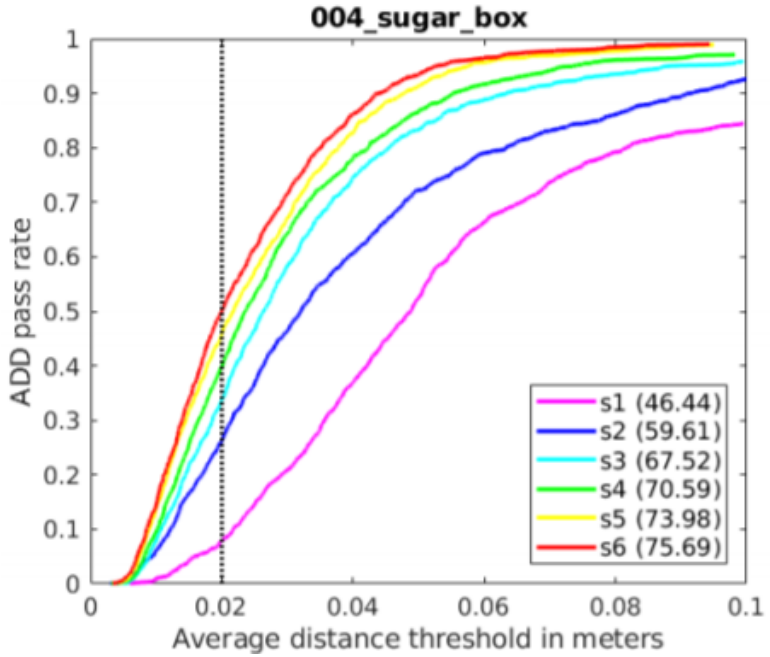


DOPE (ours)



3. Experimental Results

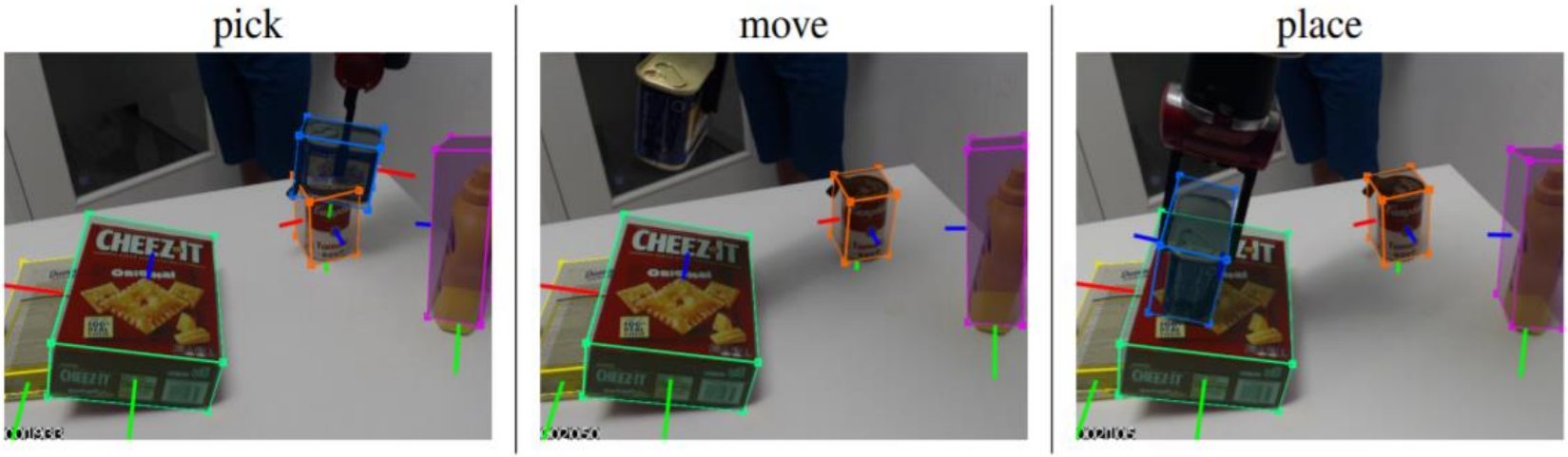
Additional experiments



	speed (ms)	AUC
1 stage	57	46.44
2 stages	88	59.61
3 stages	124	67.52
4 stages	165	70.59
5 stages	202	73.98
6 stages	232	75.69

3. Experimental Results

Additional experiments



4. Conclusion

1. We have presented a system for detecting and estimating the 6-dof pose of known objects using a novel architecture and data generation pipeline

2. We have shown that a network trained only on synthetic data can achieve SOTA performance compared with a network trained on real data, and that the resulting poses are of sufficient accuracy for robotic manipulation.

End