Articulate Object Keypoint Detection and Pose Estimation

Qi Zhu, Zhe Cao, Yaser Sheikh

Carnegie Mellon University
The Robotics Institute

Introduction

Real-time keypoint detection and pose estimation of textured objects is emerging to a fundamental problem in robotics. Among all existing detection algorithms, convolutional neural networks has been proved as the most effective algorithms. However, traditional CNN suffers from **insufficient training data**. Specifically, there are no existing dataset for articulate objects with large appearance **variation**. We developed a toolbox in Unreal Engine 4 which can automatically create large amount of high quality training images. Based on this, we explored articulate object keypoint detection and pose estimation problems.

Synthesizing training data

For single object, we set the camera in 41 different viewports and repeat the same pose sequence in each viewport. By precise keypoint projection from camera coordinate we generate images with accurate annotation.

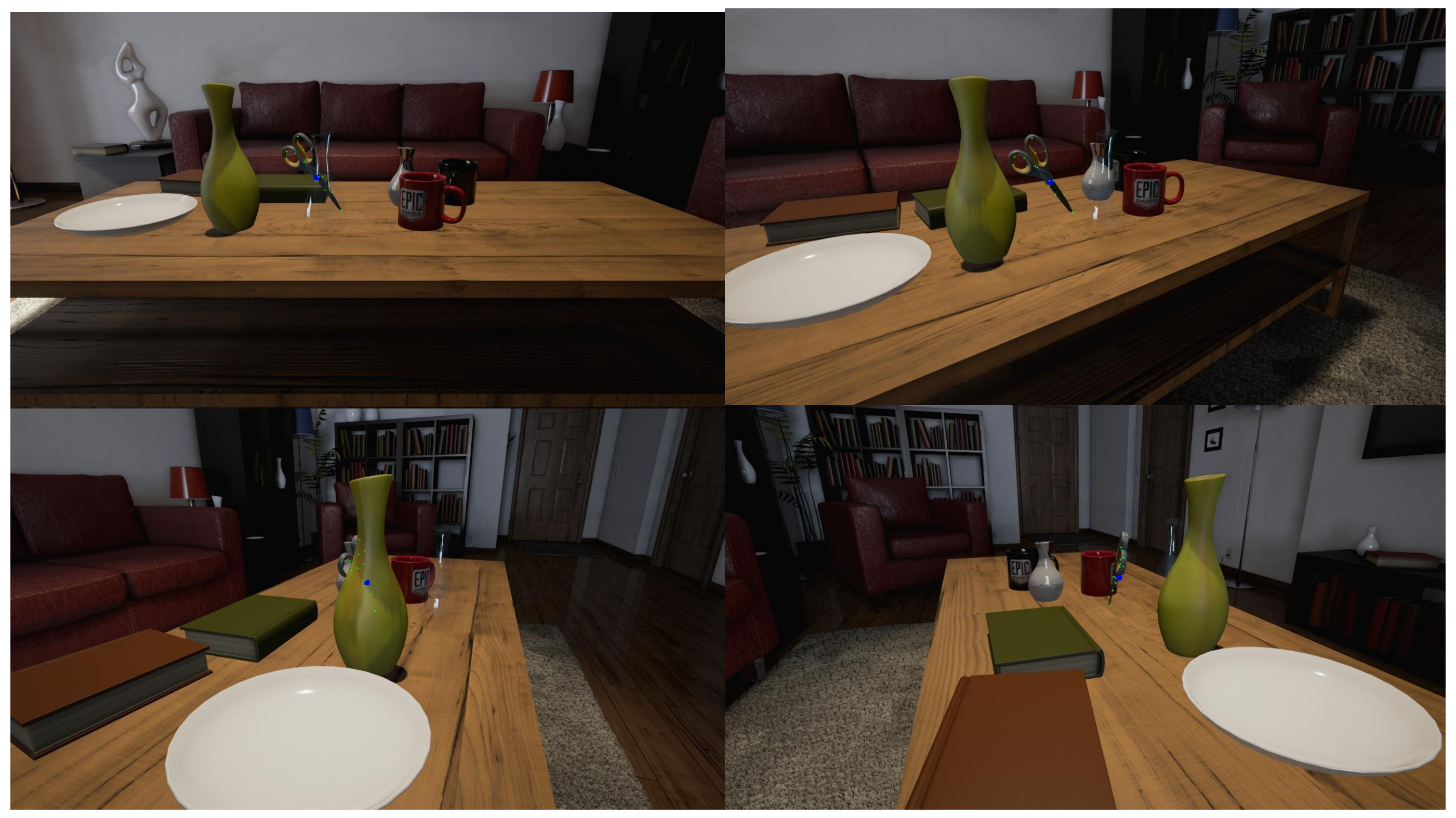

Figure 1: Keypoint annotation in different viewports

We manipulate each part of the articulate object to get each possible articulation pose and change the materials dynamically to avoid over-fitting problem.

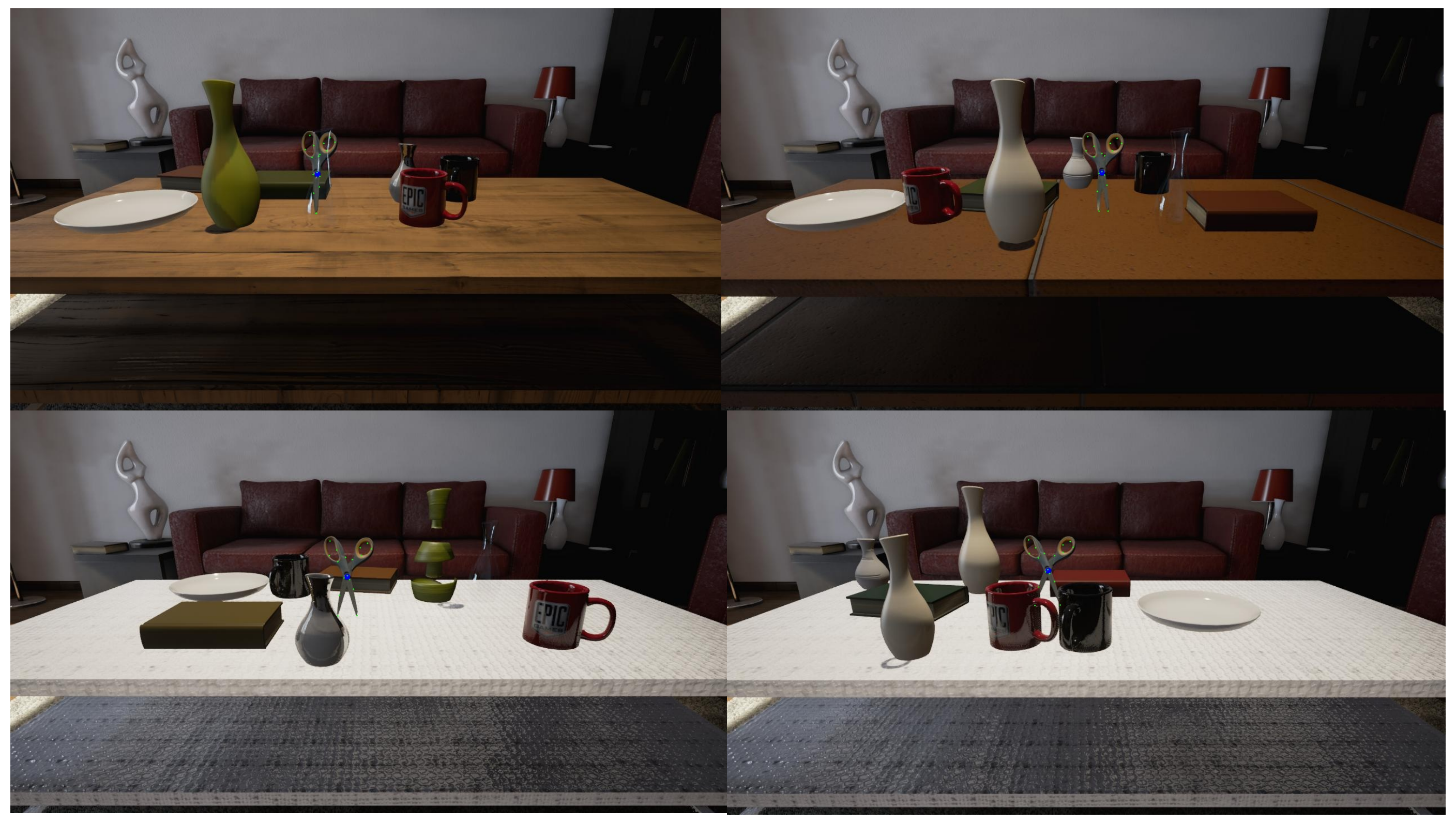

Figure 2: Keypoint annotation with different pose

Keypoint detection: Pose Machine

Convolutional Pose Machines provide a network sequence for leaning rich implicit models. It’s a deep learning architecture learn both image features and image-dependent spatial models for structured prediction task.

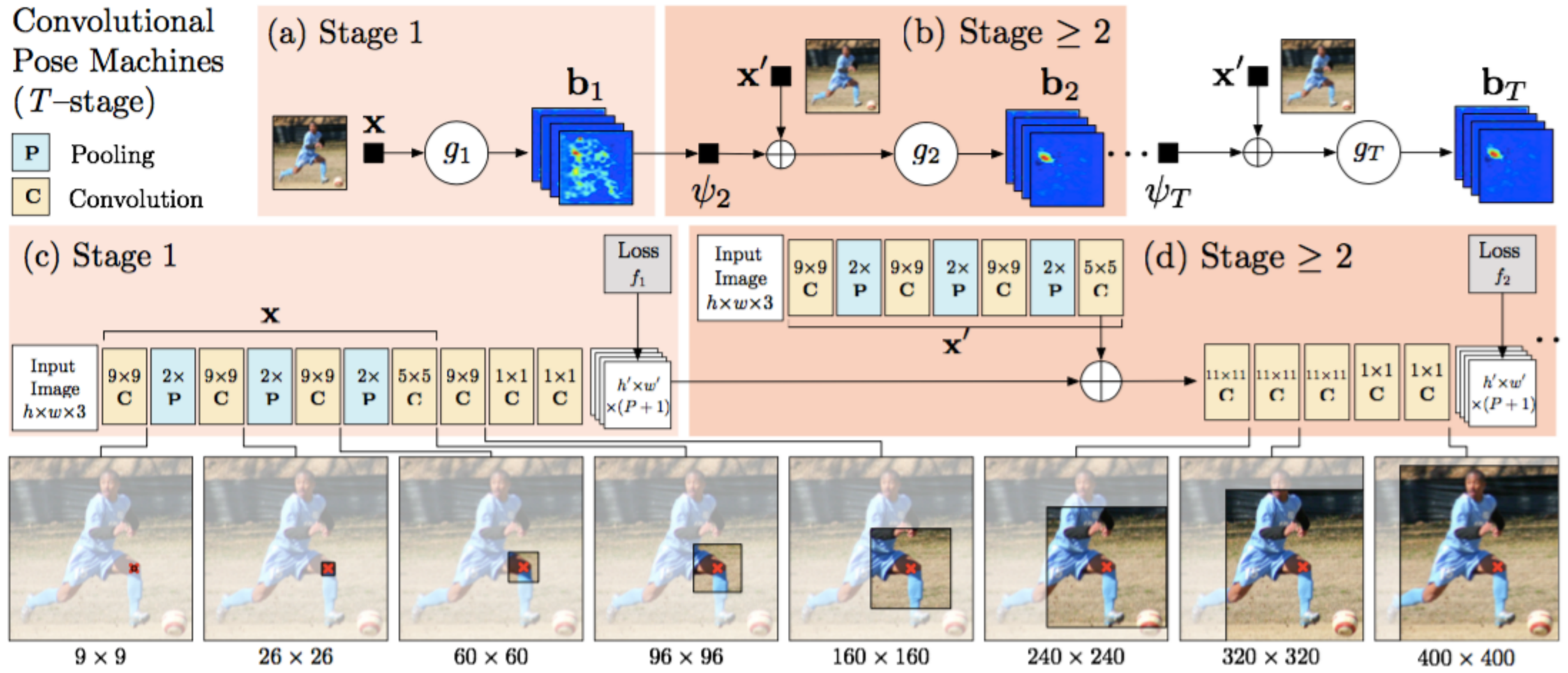

Figure 3: Architecture and receptive fields of CPMs

Articulate objects are similar with human pose. For example, to distinguish between left and right side of the scissor, context information is the best clue.

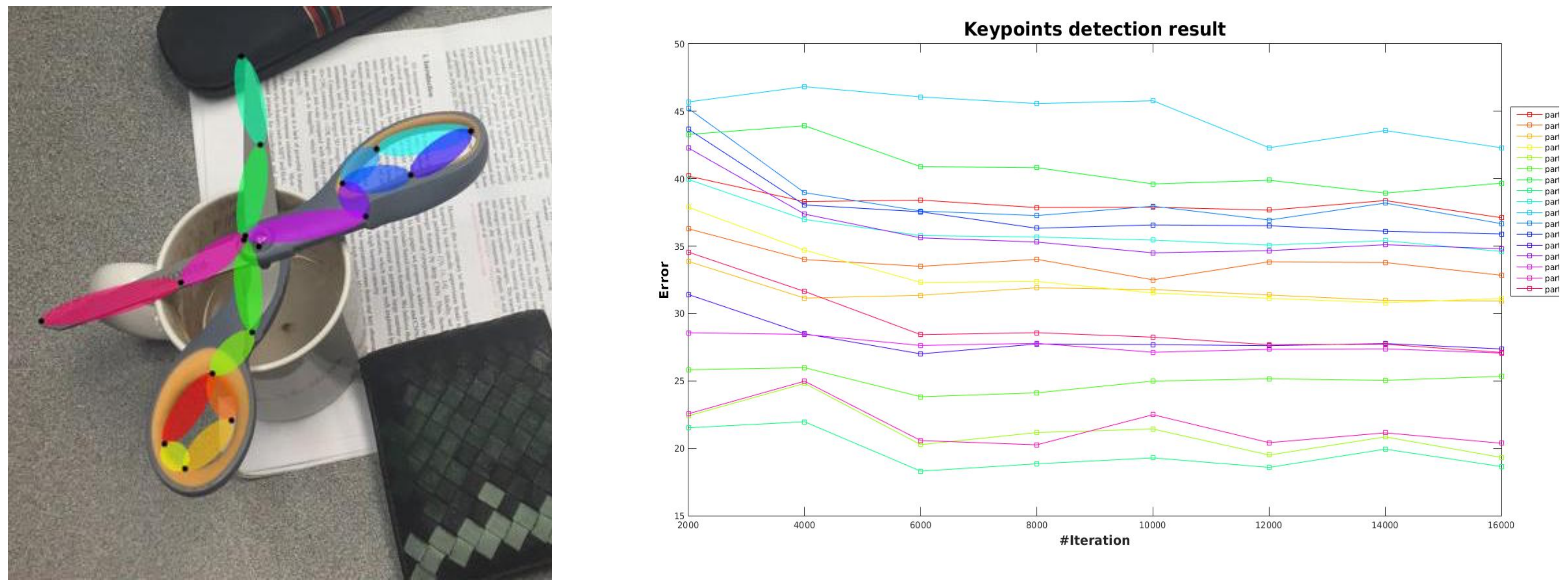

Figure 4: Qualitative result(left) and quantitative result of scissors keypoint detection

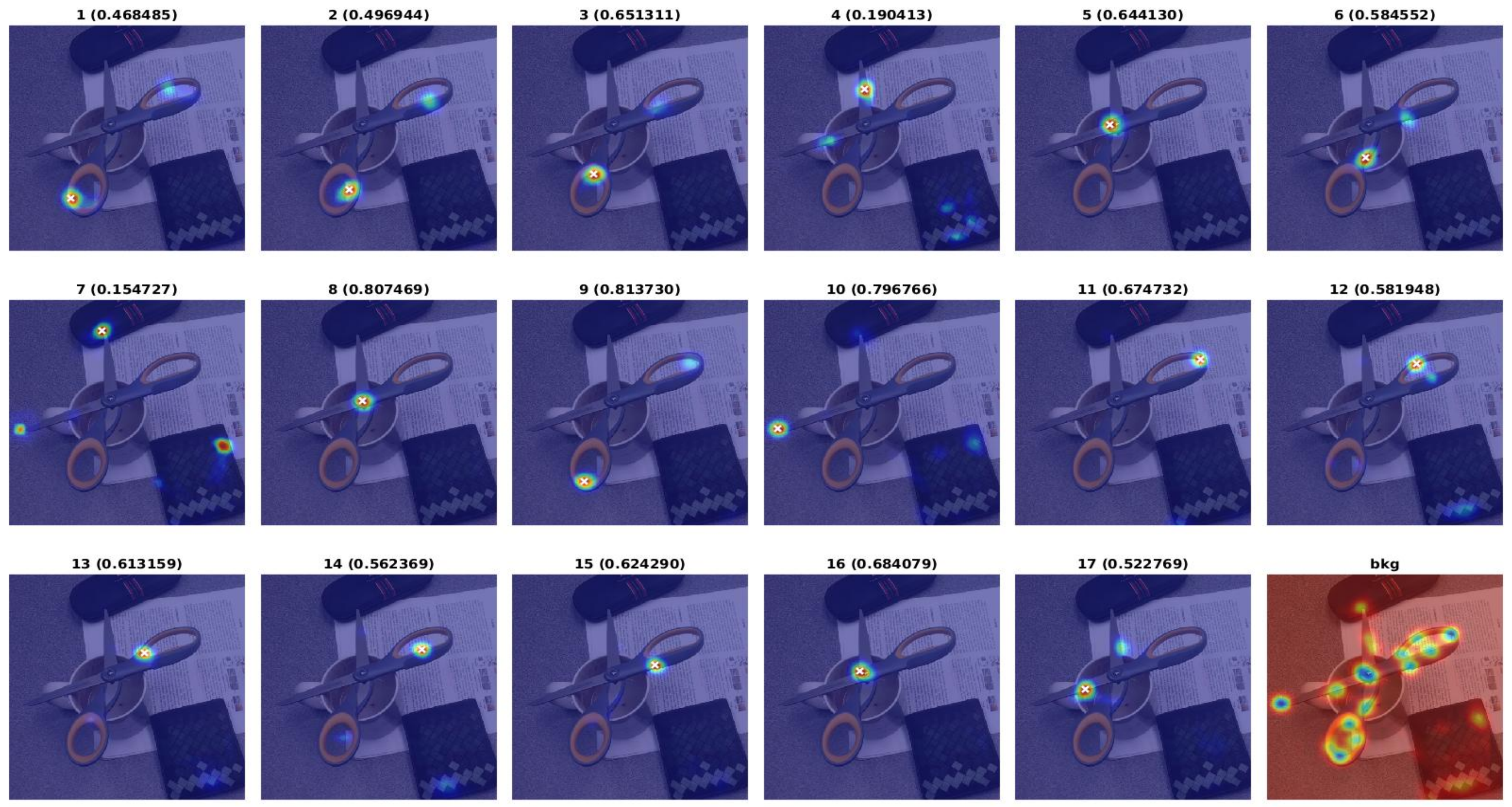

Figure 5: Belief maps for different parts

Pose estimation: PnP Algorithm

Perspective-n-Point is the problem of estimating the pose of a calibrated camera given a set of n 3D points in the world and their corresponding 2D projections in the image.

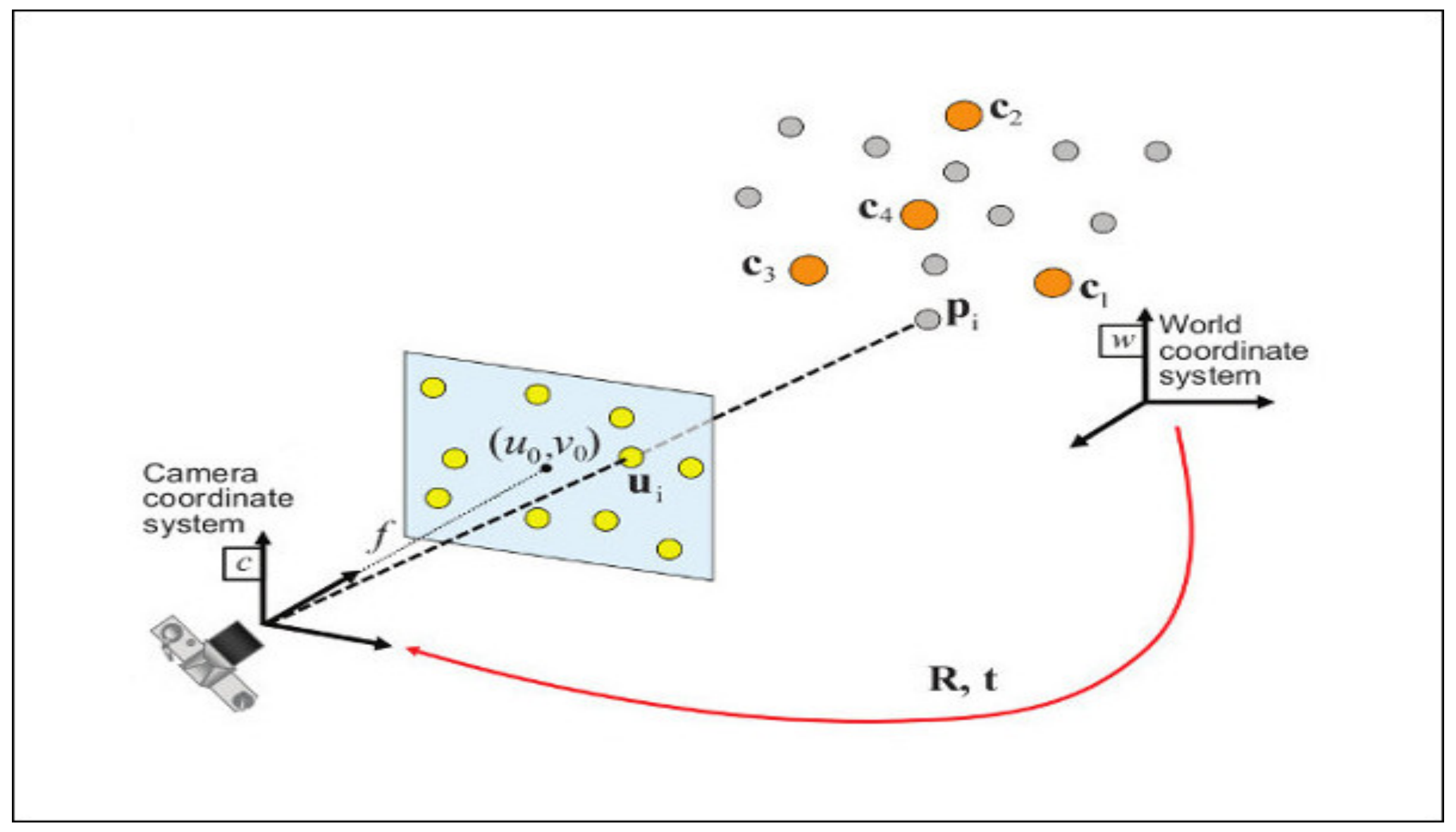

Figure 6: PnP problem illustration

$$P_{cam} = K_{proj} R_{cam} (R_w (R_o X_o) + T_w) + T_c$$

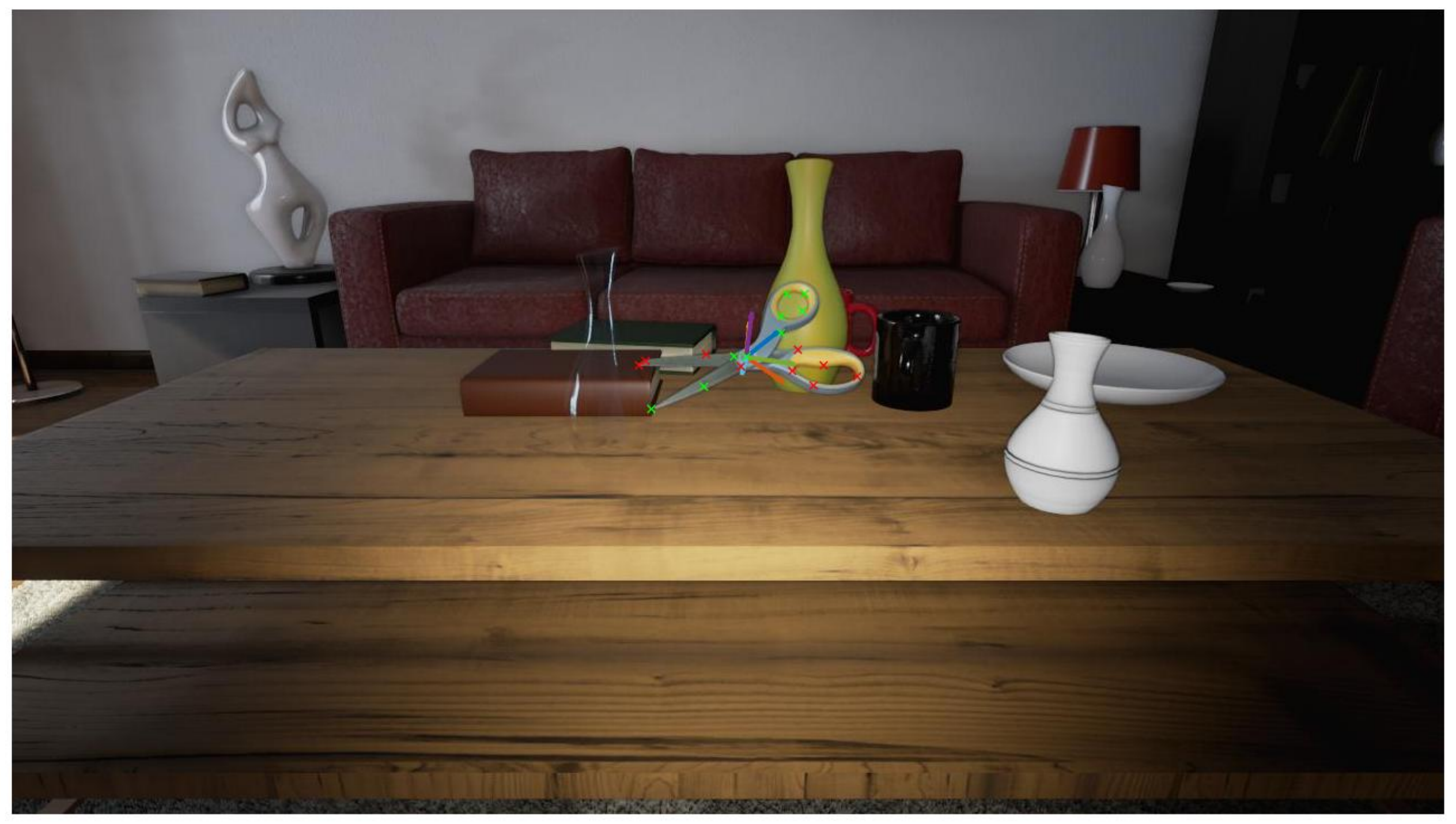

Figure 7: Pose estimation result

Conclusion

Our algorithm shows promising result in scissors. Compared with other keypoint detection or pose estimation algorithms, we are using only synthetic data. Ultimately, it will help robots identify and grasp those tricky objects in indoor scenes.

Acknowledgement

This work is funded by Toyota Motor Engineering & Manufacturing North America (TEMA, TOC, PE, PEPD, TPARC). Thanks for the advice from Zhe, Hanbyul and Thomas this summer.