

Learning Human-Object Interactions by Graph Parsing Neural Networks

Siyuan Qi^{1,2}[0000-0002-4070-733X], Wenguan Wang^{1,3}[0000-0002-0802-9567],
Baoyong Jia^{1,4}, Jianbing Shen^{†3,5}, and Song-Chun Zhu^{1,2}

¹ University of California, Los Angeles

² International Center for AI and Robot Autonomy (CARA)

³ Beijing Institute of Technology

⁴ Peking University

⁵ Inception Institute of Artificial Intelligence

syqi@cs.ucla.edu wenguanwang.ai@gmail.com baoyongjia@ucla.edu
shenjianbing@bit.edu.cn sczhu@stat.ucla.edu

Abstract. This paper addresses the task of detecting and recognizing human-object interactions (HOI) in images and videos. We introduce the Graph Parsing Neural Network (GPNN), a framework that incorporates structural knowledge while being differentiable end-to-end. For a given scene, GPNN infers a parse graph that includes i) the HOI graph structure represented by an adjacency matrix, and ii) the node labels. Within a message passing inference framework, GPNN iteratively computes the adjacency matrices and node labels. We extensively evaluate our model on three HOI detection benchmarks on images and videos: HICO-DET, V-COCO, and CAD-120 datasets. Our approach significantly outperforms state-of-art methods, verifying that GPNN is scalable to large datasets and applies to spatial-temporal settings.

Keywords: Human-Object Interaction · Message Passing · Graph Parsing · Neural Networks

1 Introduction

The task of human-object interaction (HOI) understanding aims to infer the relationships between human and objects, such as “riding a bike” or “washing a bike”. Beyond traditional visual recognition of individual instances, e.g., human pose estimation, action recognition, and object detection, recognizing HOIs requires a deeper semantic understanding of image contents. Recently, deep neural networks (DNNs) have shown impressive progress on above individual tasks of instance recognition, while relatively few methods [1, 2, 14, 38] were proposed for HOI recognition. This is mainly because it requires reasoning beyond perception, by integrating information from human, objects, and their complex relationships.

Equal contribution. [†] Corresponding author.

