VIETNAM NATIONAL UNIVERSITY - HCM CITY
UNIVERSITY OF INFORMATION TECHNOLOGY

PROJECT REPORT

COMPUTATIONAL THINKING

Class: CS117.M21

HUMAN-OBJECT INTERACTION DETECTION

Lecturer: PhD. Ngo Duc Thanh

PROJECT GROUP MEMBERS:

1.	Le Viet Thinh	20520781
2.	Dinh Nhat Minh	20521597
3.	Doan Phuong Khanh	20521443
4.	Le Nguyen Minh Huy	20521394
5.	Van Nguyen Ngoc Huyen	20521424

Ho Chi Minh City – 06/2022

CONTENT

Chapter 1. PROBLEM IDENTIFICATION

1.1. Problem Introduction

Detailed semantic understanding of image contents, beyond instance-level recognition, is one of the fundamental problems in computer vision. Detecting human-object interaction (HOI) is a class of visual relationship detection where the task is to not only localize both a human and an object but also infer the relationship between them, such as “eating an apple” or “driving a car”. The problem is challenging since an image may contain multiple humans performing the same interaction, same human simultaneous interacting with multiple objects (“sit on a couch and type on laptop”), multiple humans sharing the same interaction and object (“throw and catch ball”), or fine-grained interactions (“walk horse”, “feed horse” and “jump horse”). These complex and diverse interaction scenarios impose significant challenges when designing an HOI detection solution.

1.2. Problem Description

Understanding interactions between humans and objects is one of the fundamental problems in visual classification and an essential step towards detailed scene understanding. Human-object interaction (HOI) detection strives to localize both the human and an object as well as the identification of complex interactions between them.

Figure 1. Illustration of the output of HOT problems.
(1.1) *Interactions with similar spatial layouts can be resolved through detailed spatial information;* ***(1.2)*** *Global and local contexts encode scenes and other local objects to provide strong clues to ongoing interactions;* ***(1.3)*** *Reasonable motion estimation distinguishes between interactions where dynamics play an important role.*

The input of HOI is an image with one or more people who is doing some actions with object, then receiving a bounding box of person, a bounding box of object and a label of interaction they invovle in. The output of this problem is a pair of human-object bounding boxes and a formatted line of text that describe the interaction between human and object detected. (the annotations follow this format: <human, verb, object>)

Chapter 2. OUR SOLUTION

Human-Object Interaction problem is broken down into two stages: (1) Human Object Interaction Classification; (2) Human Object Pairs Prediction.

2.1. Phase 1 - Human Object Pairs Prediction

Prediction of a pair of person-object bounding boxes: Estimating people and objects in an image, where we solely predict things of interest (in the HOI categories under consideration) and pair each person and object.

- **Human-Detection:** Utilizing Deep Learning Applications (we apply the Faster-RCNN method proposed in *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks* paper to solve this subproblem).
- **Object-Detection (for each HOI category):** Considering each HOI category, only predict the object of interest in that category. Utilizing Deep Learning Applications (we apply the Faster-RCNN method proposed in *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks* paper to solve this subproblem).

2.2. Phase 2 - Human Object Interaction Classification

- **Examining Human-Object spatial relations through Pattern Recognition**

Splitting into 3 different streams to extract the features of each separate stream (Human, Object, Action)

- **Human Stream:** Extracting local features from the human bounding box and generating confidence scores for each HOI class. Utilizing ConvNet (we apply the VGG16 method proposed in *VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION* paper to solve this subproblem).
- **Object Stream:** Extracting local features from the object bounding box and generating confidence scores for each HOI class. Utilizing ConvNet (we apply the VGG16 method proposed in *VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION* paper to solve this subproblem).

- **Pairwise Stream:** Extracting features that encode the spatial relations between the human and object, generating a confidence score for each HOI class.

DNNs can learn 2D filters that respond to similar 2D patterns of human-object spatial configurations. We would use a special type of DNN input that characterizes the relative location of two bounding boxes (whose Interaction Pattern is a binary image with two channels: The first channel has the value of 1 at pixels enclosed by the first bounding box, and the value of 0 elsewhere; the second channel has value 1 at pixels enclosed by the second bounding box, and value 0 elsewhere.). The first channel corresponds to the human bounding box and the second channel corresponds to the object bounding box. To get the Interaction Pattern we would use an Attention window minimized as small as possible that contains two bounding boxes.

- **Action Classification:**

The action classification should be considered as a multi-label classification as opposed to the traditional K-way classification. The classification model is trained by applying a sigmoid entropy loss on the classification output of each class, then computing the total loss by summing all the individual losses.

- **Element-wise Sum:** The final score is summed over all streams separately for each HOI class.
- **Selecting:** Considering confident scores on all HOI classes of human-object pairs, choose the label of the image according to the criterion of highest confidence score.

Chapter 3. THE APPLICATION OF COMPUTATIONAL THINKING

3.1. Decomposition

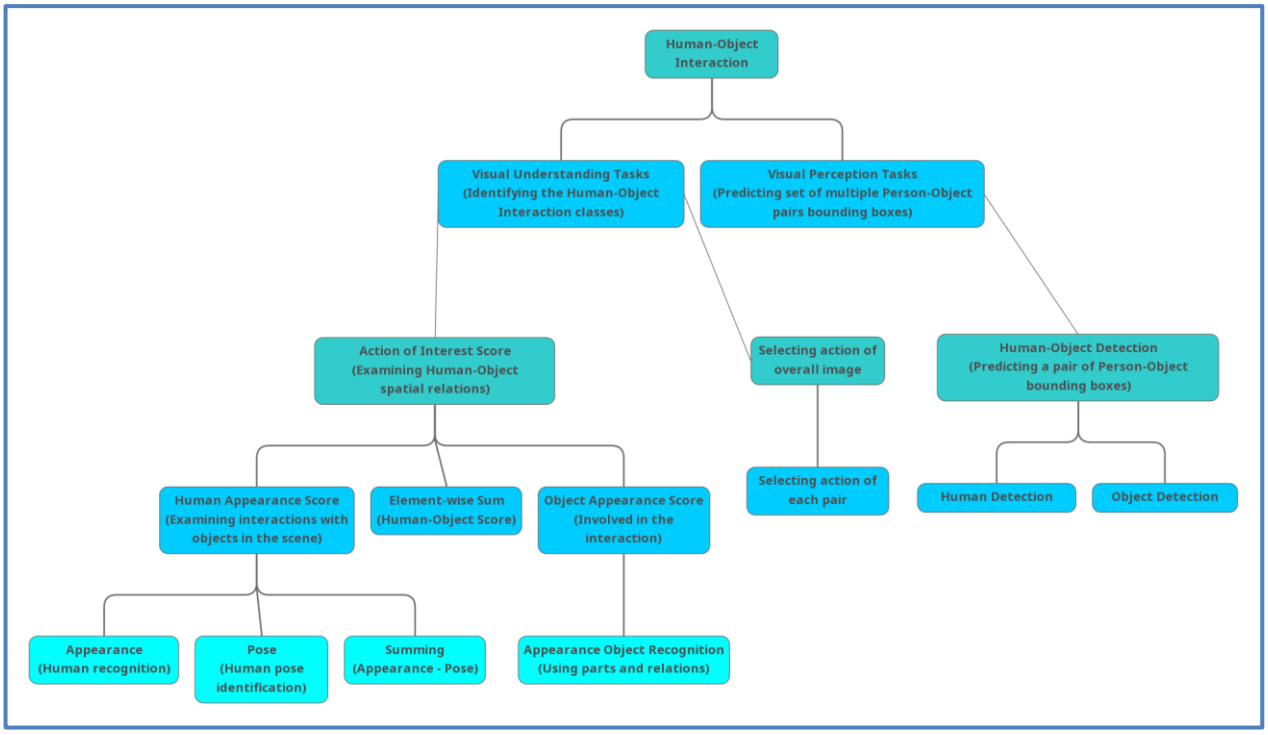

Figure 2. The Decomposition Tree of the Human-Object interaction problem.

THE INPUT AND OUTPUT OF EACH UNIT PROBLEM

Human-Object Interaction

Input: A picture that has a horizontal side view of human(s) interacting with random objects.

*Output:*1. Minimum-sized rectangle bounding boxes of human(s) and object(s).

2. A formatted line of text that describe the interaction between human and object detected. (the annotations follow this format: <human, verb, object>)

First Branch: Visual Perception Tasks

Input: A picture that has a horizontal side view of human(s) interacting with random objects.

Output: Multiple human-object pair with minimum-sized bounding boxes.

- ✓ **Human - Object Detection**

Input: Minimum-sized rectangle bounding boxes of each human and object in the image.

Output: A human-object pair.

- ○ **Human Detection**

Input: A picture has a horizontal side view of human interacting with random objects.

Output: Minimum-sized rectangle bounding boxes of each human in the image.

Figure 3. Human detected in bounding boxes

- **Object Detection**

  Input: A picture has a horizontal side view of human interacting with random objects.

  Output: Minimum-sized rectangle bounding boxes of each object in the image.

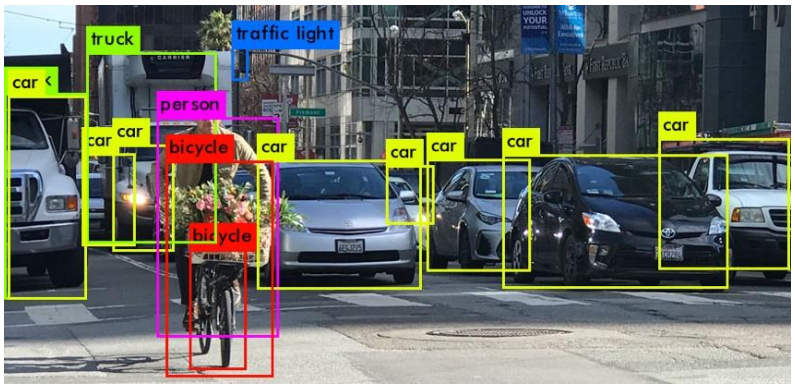

Figure 4. An Object-Detection result in side-view image

Second Branch: Visual Understanding Tasks

Input: Multiple human- object pairs with minimum-sized bounding boxes.

Output: A formatted line of text that describe the interaction between human and object detected. (the annotations follow this format: <human, verb, object>)

✓ **Actions of Interest Score**

Input: A human-object pair with bounding box and a list of actions of interest.

Output: A vector containing scores of all listed actions.

- **Human Appearance Score**

  Input: A human-object pair with bounding boxes and a list of HOI classes of interest.

  Output: A vector containing scores of all listed HOI classes of interest in which the human could be involved.

  - **Appearance**

    Input: A list of objects of interest in which the human could be involved and bounding box of the human.

    Output: A vector containing scores of all listed objects of interest in which the human could be involved.

- **Pose**

Input: A list of action of interest in which the human could be involved and bounding box of the human.

Output: A vector containing scores of all listed actions of interest in which the human could be involved.

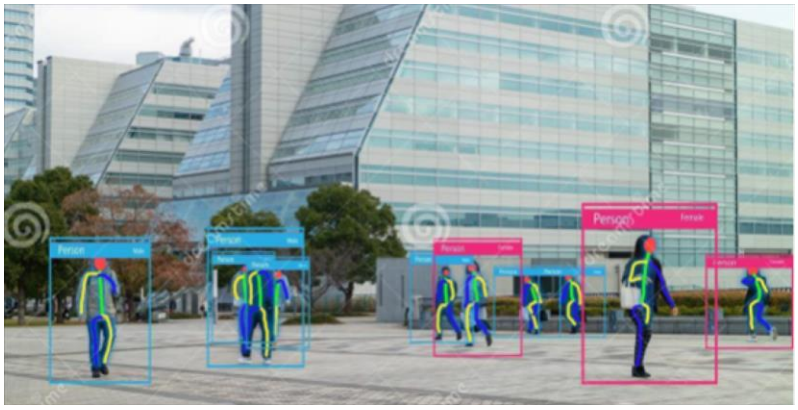

Figure 5. Human-pose recognition

- **Summing (Appearance - Pose)**

Input: (1) A vector containing scores of all listed objects of interest in which the human could be involved.

(2) A vector containing scores of all listed actions of interest in which the human could be involved.

Output: A vector containing scores of all listed HOI classes of interest in which the human could be involved.

Figure 6. Human's action of interest scoring

- **Element-wise Sum (Human-Object Score)**

Input: (1) A vector containing scores of all listed actions in which the human could be involved.

(2) A vector containing scores of all listed actions in which the object could be involved.

Output: A vector containing scores of all listed actions in which the object and the human could be involved.

- **Object Appearance Score**

  Input: A human-object pair with bounding boxes and a list of actions of interest.

  Output: A vector containing scores of all listed actions that object could be involved.

✓ **Selecting Action of the overall image**

Input: A vector containing scores of all listed actions of a human-object pair.

Output: (1) Label of maximum scored action out of all actions.
(2) Score of the selected action.

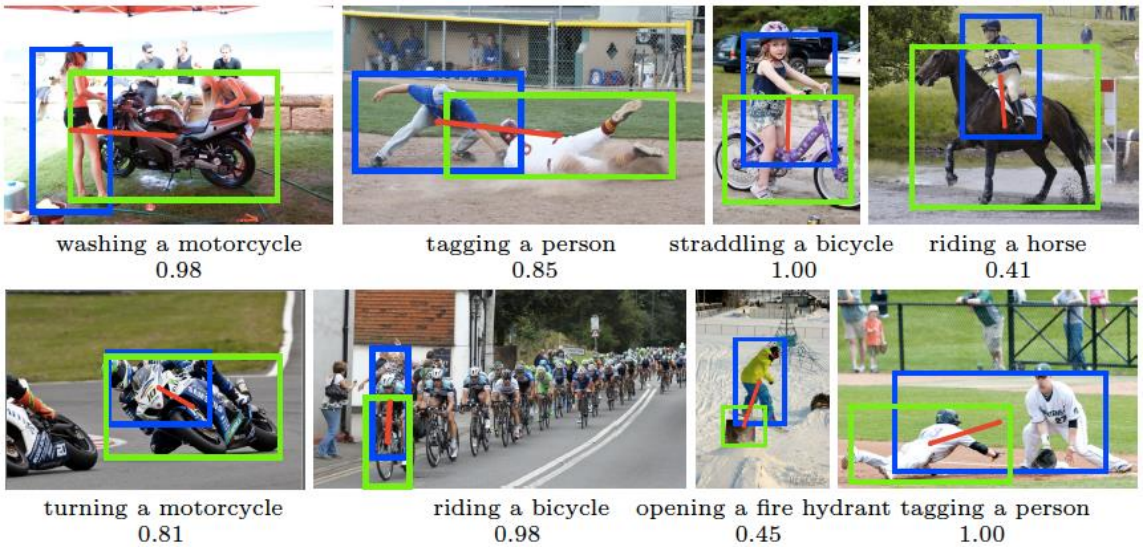

Figure 7. Action's Selection and Scoring

3.2. Pattern Recognition

Patterns within the **Human-Object Interaction Problem** exist within its smaller problems that we have decomposed to.

In the problem of Human-Object Detection, it's 2 smaller problems: Human Detection and Object Detection have the same pattern as the Detection problem. So we could group them as a single problem: Detecting existences within images.

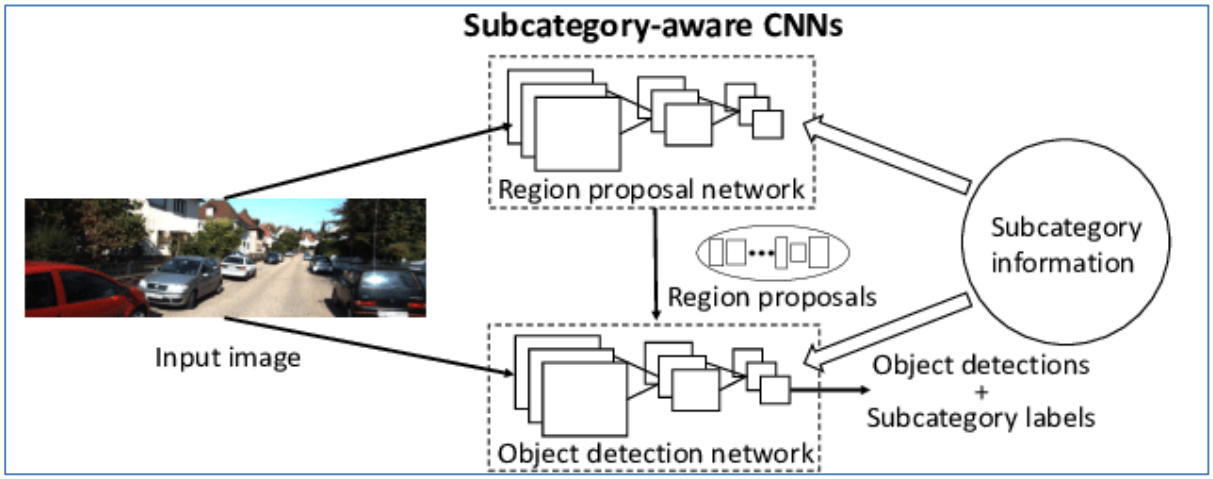

Figure 8. Deeplearning-based Object detection framework

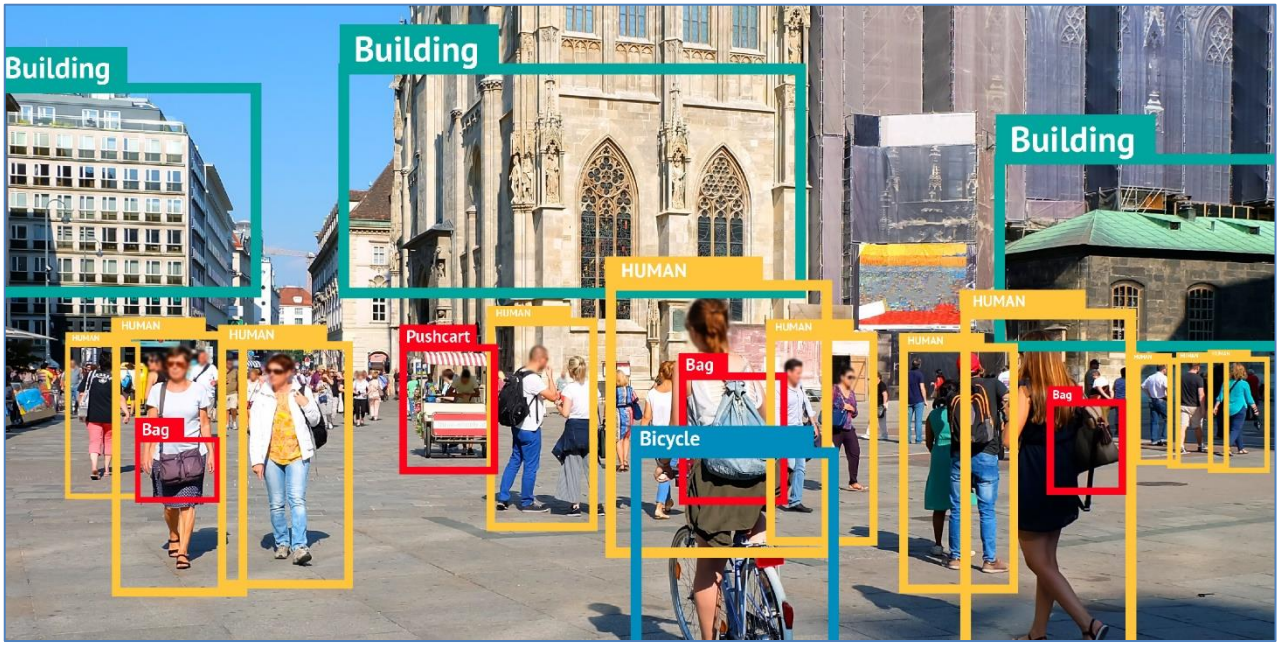

Figure 9. The Humans and Objects localization indicated by Object Detectors

In the problem of Scoring interested action through examining human-object spatial relations, its two smaller problems:

- ✓ Scoring the Human appearance involved in the interactions with objects in the scene
- ✓ Scoring the appearance of the object involved in the interaction

The above problems share the same pattern as their common requirement is characterizing the relative location of two bounding boxes.

The Selecting action of the overall image appeared to have the pattern of the Multiclass Classification problem- a central topic in machine learning that has to do with teaching machines how to group data together by particular criteria.

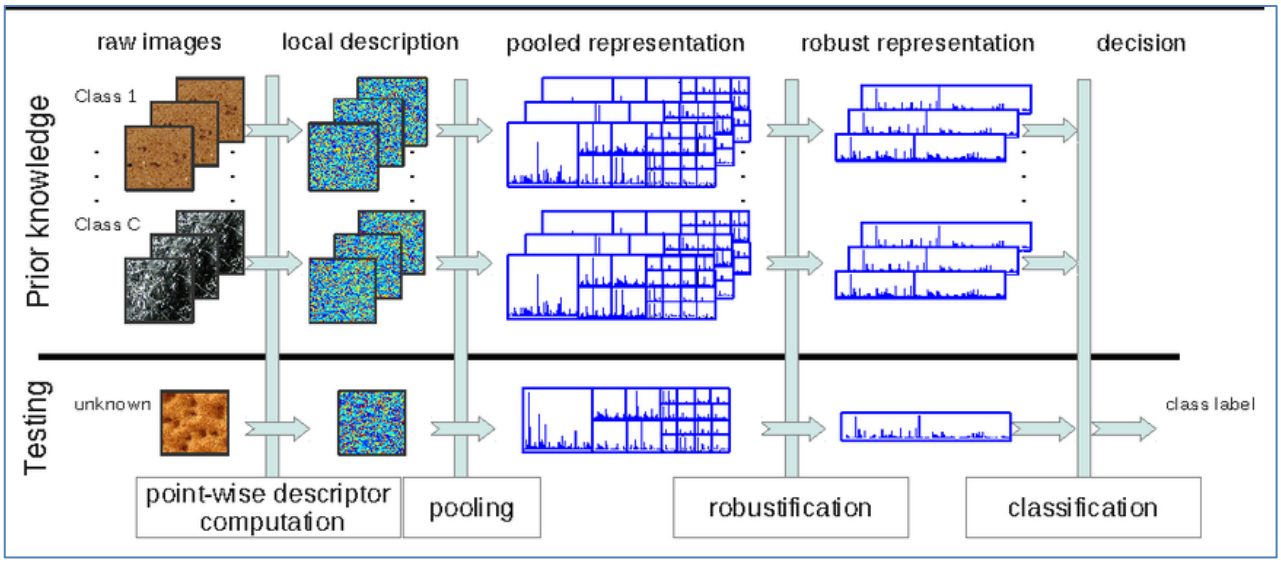

Figure 10. Deeplearning-based Classificaton framework.

3.3. Abstraction

The inputs and outputs of the Human-Object Interaction problem as well as its subproblems have been elaborately described in Section 3.1 For more specific:

The Human and Object Detection tasks are grouped as the sole detection tasks. So we don't need to take into account whether its input contains specific things or humans, and the output may just contain all the existence within the image. With that in mind, we may leverage a significant number of object detection models that are well-constructed and easily implemented to address this subproblem. So that, in the input, we may take A picture with no more specific details about the humans or object included in the image. The output of the deep learning model would give us bounding boxes of multiple humans and objects.

The Selecting Action of overall picture tasks that we have mapped into the Multiclass Classification problem might also be applied to the Deep learning-based classification task, in which a considerable number of models have been suggested and developed over decades. With the use of Deep learning, the Scoring Action of Interest problem could be grouped and mapped into classification tasks as the output of classification models could be a vector containing scores of all listed labels as well.

Chapter 4. CONCLUSION

Accurate recognition of Human-Object Interaction can benefit numerous tasks in computer vision, such as action-specific image retrieval, caption synthesis, and question answering. Leveraging from earlier achievements in visual recognition, in this work we focus on Human-Object Interactions for machines to comprehend what is happening in visual photographs. Thanks to the Computational Thinking process, after decomposing this complex and multifarious problem into a simple and manageable one, we propose a solution based on the multi-stream approach to recognize interactions in each instance by utilizing object detection and classification framework. We hope that our technique could provide computers with a richer knowledge of the semantics of visual images as well as potentially affect the method design procedure in this field of research.

Chapter 5. REFERENCE

[1] T. Bergstrom and H. Shi, "Human-Object Interaction Detection: A Quick Survey and Examination of Methods", Proceedings of the 1st International Workshop on Human-centric Multimedia Analysis, pp. 63-71, Oct. 2020.

[2] Wang, T., Yang, T., Danelljan, M., Khan, F.S., Zhang, X., Sun, J.: Learning human-object interaction detection using interaction points. Proc. IEEE Conf. Computer Vision and Pattern Recognition (2020).

[3] Yao B, Fei-Fei L (2010) Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR

[4] Maraghi, V. O., & Faez, K. (2021b). Scaling Human-Object Interaction Recognition in the Video through Zero-Shot Learning. Computational Intelligence and Neuroscience, 2021, 1–15. https://doi.org/10.1155/2021/9922697

[5] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. Deep contextual attention for humanobject interaction detection. In ICCV, 2019

[6] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng. Learning to detect human - object interactions. In arXiv preprint arXiv:1702.05448, 2017

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014

[8] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015