EXPLORING HUMAN-OBJECT INTERACTION DETECTION

by

TREVOR BERGSTROM

A THESIS

Presented to the Department of Computer Science
And the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Master of Science

September 2020

THESIS APPROVAL PAGE

Student: Trevor Bergstrom

Title: Exploring Human-Object Interaction Detection

This thesis has been accepted and approved in partial fulfillment of the requirements for the Master of Science degree in the Department of Computer and Information Science by:

Humphrey Shi	Chair
Dejing Dou	Core Member

and

Kate Mondloch	Interim Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded September 2020

THESIS ABSTRACT

Trevor Bergstrom

Master of Science

Department of Computer and Information Science

August 2020

Title: Understanding Human Object Interaction Detection

Human-object interaction detection is a relatively new task in the world of computer vision and visual semantic information extraction. The goal of human-object interaction detection is to have machines identifying interactions that humans perform on objects. We provide a basic survey of the developments in the field of human object interaction detection. Many works in this field use multi-stream convolutional neural network architectures, which combine features from multiple sources in the input image. To provide insight to future researchers, we perform a study examining the performance of each component of a multi-stream architecture for human-object interaction detection. We examine the HORCNN architecture as a foundational work in the field. We also provide an in-depth look at the HICO-DET dataset, a popular benchmark in the field of human object interaction detection. Lastly, we begin the construction of a human-object interaction benchmarking platform.

CURRICULUM VITAE

NAME OF AUTHOR: Trevor Bergstrom

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
Central Washington University, Ellensburg
Oregon State University, Corvallis

DEGREES AWARDED:

Master of Science, Computer and Information Science, 2020, University of Oregon
Bachelor of Science, Mechanical Engineering Technology, 2014, Central Washington University

AREAS OF SPECIAL INTEREST:

Machine Learning
Computer Vision

PROFESSIONAL EXPERIENCE:

Software Engineering Intern, Insitu inc, 2020

R&D Software Engineering Intern, Moovel LLC, 2019

Graduate Teaching Fellow, University of Oregon, 2018 - 2020

Manufacturing Project Engineer, Genie Industries, 2015 - 2018

Composites Design Engineer, Amtech LLC, 2014 – 2015

PUBLICATIONS:

Bergstrom, T. Shi, H., "Human-Object Interaction Detection: A Quick Survey and Examination of Methods," In Proceedings of ACM Multimedia 2020: 1st International Workshop on Human-centric Multimedia Analysis, 2020.

ACKNOWLEDGMENTS

I would like to thank all the faculty in the Department of Computer and Information Science at the University of Oregon for supporting me through this degree. I would particularly like to thanks Dr. Humphrey Shi for inspiring me and guiding me through the process of learning how to become an academic researcher. Lastly, I could not have made it through without the support of Kyeti Morgan and my parents, Aileen and Bob Bergstrom.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

Achieving the goal of true machine intelligence requires an agent that can observe and understand its environment just as humans are able to. There has been a significant amount of excitement and progress around machine learning and its ability to solve problems related to emulating human understanding of our natural and social environments. The field of computer vision, in particular, has recently exploded with the advent of deep learning techniques that can solve complex object detection problems. However, simply identifying objects in an image is not what should be considered *true machine intelligence*. Striving towards the idea of more intelligent machines, researchers have created models and systems that can extract richer semantic information from images and videos. As humans, we are able to recognize relationships between objects in an image. These relationships can help an intelligent machine interpret the underlying meaning of the image or scene, and therefore, take one step closer to understanding the world around us.

Visual scene understanding is a complex set of interpretations about what is happening in an image. Full understanding of a scene can be separated into two classes of understanding, perception, and context reasoning. Perception is defined as organization, identification, and interpretation of sensory information. Reasoning can be defined as the capacity of consciously making sense of things, applying logic, and adapting or justifying practices and beliefs based on new or existing information. Perception tasks in visual understanding include object detection and visual semantic segmentation. These tasks can be seen as observing and identifying visual information, though little reasoning is used to

help accomplish them. Context reasoning applications include visual relationship detection, visual questioning answering, and scene graph generation. These tasks seek a deeper understanding of what is happening in an image, using reasoning to interpret the harder to detect visual information that the image contains. It should be said that the latter task cannot be accomplished without the former. In other words, humans usually perceive first, then reason. Observe then interpret.

Commonly, when humans seek to interpret their environment, they do so by observing other humans and how they interact with one another or objects. Object to object interactions, for the most part, deal with simple spatial or descriptive interactions such as *book on top of table* or *chair on floor*. Humans can provide a much richer set of interactions with objects, as there are visual and non-visual ways a human can interact with their natural environment. A benefit of examining humans is in the unique ways we display intent and interact. The appearance of a human performing an action can be viewed through fine-grained attributes such as body positioning and placement, or even gaze. All of these attributes give deeper and richer semantics that we can use to identify human actions. This work will focus primarily on the task of *human-object interaction detection*. The goal of human-object interaction detection (HOI), is to correctly identify humans, objects, and the actions that are occurring between them, if any, in an image. These action relationships are commonly represented in triplet form, {human, action, object}. The first step in discovering an HOI from an image is to detect objects. Object proposals recovered from the image should contain at least one human for an HOI to be present. Using these object proposals, a model for solving this problem must then

correctly identify an HOI between the humans present and any of the objects in the image.

1.1 Proposed Work

Human-object interaction detection is a relatively new field in the computer vision research community. Many models provide a good starting point but do not perform as well as many algorithms in other fields of computer vision. A foundational labeled dataset for this task is the HICO-DET [6] dataset created by Chao et al in 2018. HICO-DET has become a standard in benchmarking human-object interaction detection models, providing two separate data arrangements for evaluation. However, the dataset and its evaluation metrics are not easily compatible with modern deep learning frameworks.

The first contribution of this work is a toolkit of software for human-object interaction detection, similar to that of mmdetection [8] which is used for object detection. The first major component of this is to address the dataset. By creating a data loader that seamlessly integrates with the PyTorch deep learning framework [46]. Secondly, since the first step of human-object interaction detection is object detection, a highly accurate and robust object detector is required in a detection pipeline. The data loader can be easily integrated with detectors providing proposals for human and objects, taken from images, in bounding box coordinate format. This integration will also pre-compute these proposals for training rather than performing detection during training, freeing up GPU, and system memory for the main components of the detection pipeline. These components build the foundation for what we consider a toolkit for HOI detection tasks. We provide the user with pre-trained models, as well as training and testing

infrastructure for future models, with seamless dataset integration. We hope that this work can provide future researchers with a single platform, flexible toolkit to further develop and improve the field of human-object interaction detection.

The second contribution proposed in this work is an in-depth study of the performance of individual model components. Since a majority of the state-of-the-art models use a multi-stream network, we want to examine the individual performance of each component separately. This investigation can provide useful information on how to develop future models, using this multi-stream method. Using the components from HORCNN [6], we conduct numerous tests on their ability to correctly classify human-object interactions.

1.2 Overview

Chapter 2 of this work covers background information starting with an overview of machine learning techniques for computer vision tasks and including an overview of related work and datasets in human-object interaction detection. Chapter 3 provides in-depth detail on building the data loader for the HICO-DET dataset, as well as our implementation of the HORCNN model [6]. Chapter 4 discusses the results obtained from our study into the separate components of the HORCNN model, chapter 5 discusses future work on this toolkit, and improvements to current and future models and datasets for the task of human-object interaction detection. Finally, chapter 6 is where we will draw our conclusions.

CHAPTER II

BACKGROUND AND RELATED WORKS

Human object detection is closely related to other fields in computer vision. Many of the current state-of-the-art human-object interaction models draw inspiration from these methods. This section will begin by introducing the building blocks for deep learning and computer vision. An overview of related tasks in computer vision will follow.

2.1 Background

For an intelligent agent to begin to understand its environment it must first make observations, gathering data to process. One of the first perceptions of an environment humans make is to look and make visual observations. This may be a simple task for humans, but it poses numerous complex problems for a computer to replicate this process. The field of machine vision, or more commonly computer vision, is a subfield of artificial intelligence in which researchers and engineers seek to teach a computer to make complex visual observations. The first step is to create a visual representation that a computer can interpret. Digital images are an example of this, consisting of stacked two-dimensional matrices of pixels that represent color intensity, known as channels, in a finite and discrete numerical representation. Each channel represents a color, typically red, green, and blue. See Figure 2.1 for an illustration of this. From the numerical representations, computers are able to use this data as input. Using this most basic form of data, it is possible to teach a computer to find interesting parts of an image, known as visual features, to begin processing this visual data. Visual features can include edges, blobs, or corners that help distinguish sections of an image. Given these most basic

building blocks of digital visual representation, computers are able to make more complex observations of their environment.

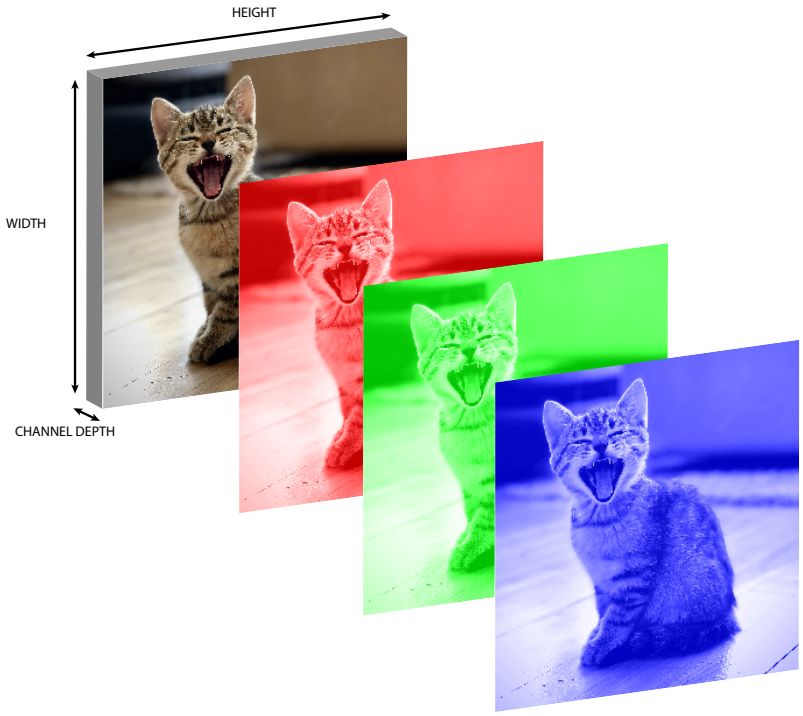

Figure 2.1: Images stored as 3D matrices.

2.1.1 Convolutional Neural Networks

Convolutional neural networks (CNNs) are frequently used for visual understanding tasks, they show improved performance over multi-level perceptron neural networks (MLPs), due to the fact that they don't suffer from the loss of spatial information while interpreting an image [69]. Intuitively, convolutional layers pass filters over two-dimensional patches of the input image to extract learned features, such as edges of an object. Conversely, a standard fully connected neural network would need to vectorize the image into a single dimension to process, losing the spatial relationships between the pixels in an image. A convolutional network usually involves a few parts.

The Convolution layer will perform the aforementioned convolutional operations, in which an output, referred to as a feature map, is generated. The feature map represents only specific features of the data that are needed for the task at hand, filtering some unnecessary information from the input. Next, a rectified linear unit (ReLU) layer follows, acting as the activation layer, this transforms the output feature map by removing negative pixel values. The last component is the pooling layer, which performs a form of non-linear down-sampling [10]. This layer scales the image size down, providing an abstracted representation of the image features which can help avoid overfitting. It also reduces the overall computational and storage costs of the network. Typically, after a series of these layers, classification is performed using fully connected layers to get the final output values of a network [39]. The breakthrough ability of convolutional networks is, through extensive training, the convolutional layers learn to generate the correct features for the dataset [26]. Before convolutional neural networks, features for many computer vision tasks needed to be painstakingly hand-engineered. This difficult process provided poor generalization across datasets and various categories of images [69]. One of the first advances in convolutional network design was that of AlexNet [32], the architecture of this network is shown in Figure 2.2 for a reference.

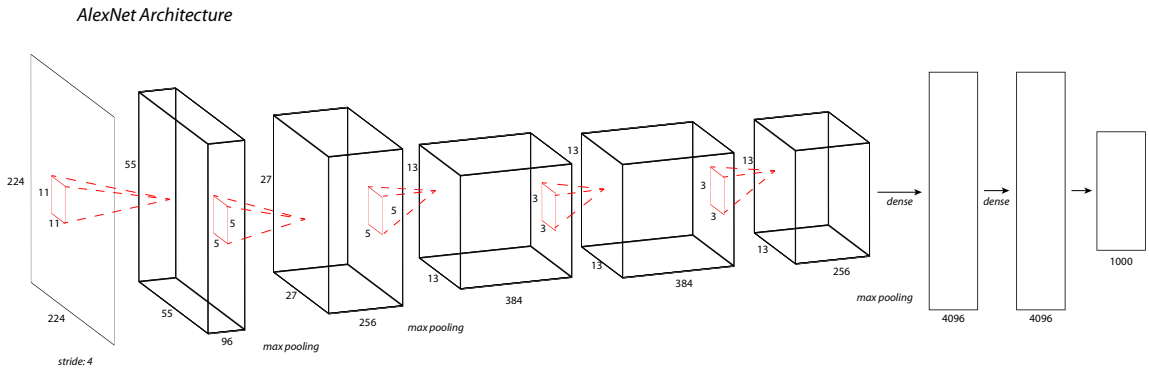

Figure 2.2: Architecture of AlexNet convolutional neural network for image classification.

2.2 Visual Perception Tasks

Visual perception is the starting point of understanding a scene or image. To begin to understand the scene, perception such as identifying the objects in the scene is imperative. The following tasks, image classification, and object detection are fundamental for more elaborate visual understanding. In this section, a basic overview of each is provided.

2.2.1 Image Classification

Image classification is a foundational task in computer vision. In image classification, the goal is to identify the object contained in the image. Most datasets for this task are easy by modern standards, usually containing just the object in question and very little background noise. A typical model for image classification contains a series of convolutional layers followed by max-pooling operations that downsample the feature

maps. Finally, a series of fully connected layers perform final object classification on the feature map [69].

2.2.2 Object Detection

The task of object detection seeks to locate and identify specific classes of objects in the image. The maturation of this field has allowed researchers to accelerate progress in other more complex areas of computer vision such as visual understanding. Human-object interaction relies heavily on its first step of object detection. With the popularization and rise of deep learning methods, models are able to perform very well on large and complex datasets [69]. Object detection can be thought of as another abstraction on top of image classification. Object detectors need to first find an object within an image, then classify that object. These two steps are commonly referred to as localization and classification. The goal of localization is to find regions of an image in which an object might exist, we call these areas region proposals. The classification step employs methods similar to image classification, where convolutional features are used to determine the class of the object.

Region proposals are defined by a bounding box, a set of spatial coordinates within the image, and a class label identifying the object. Finding the object is a complex task. One naive way of finding regions would be to sample all areas and all bounding box sizes in the image. However, this would be computationally inefficient as there could be a near-infinite number of bounding boxes to perform classification on. There has been a great deal of effort and research into the task of localization, and consequently, models handle region proposals generation differently. See Figure 2.3 for an example of localized

objects in an image. Commonly, modern object detection algorithms are divided into two classes; two-step detectors and single-shot detectors [69].

Some examples of two-step detectors are the RCNN [15] [16] [52] family of detectors. Two-step detectors, as the name implies, require two separate steps (localization and classification) to detect objects from a given image. Region proposal generation can be accomplished in a number of different ways. In the case of RCNN [16], a selective search algorithm is used, which employs pixel similarity metrics to determine possible connected pixel groupings. Faster-RCNN [52] uses a CNN-based approach called the region proposal network, which discovers region proposals from convolutional feature maps. After localization is performed, classifying these regions can be accomplished by simple feedforward neural network classifiers. Two-step detectors typically show better accuracy than their single-shot counterparts, however, the detection time is much higher [69].

Single-shot detectors operate by performing bounding box localization and object classification at the same time. Some well-known models that use this method of object detection are the Single-Shot Detector (SSD) [41] and the You Only Look Once (YOLO) family of models [51]. The SSD uses feature maps at multiple sizes. These feature maps are the outputs of the convolutional layers after each max-pooling downsampling operation. On these feature maps, multiple default bounding boxes of assorted sizes are laid over the feature map to create the region proposals. These regions are then classified using the convolutional features already existing within them. Since both classification and localization are done in a single step, single-shot detectors are very fast at prediction and can perform object detection on high-frame-rate videos. However, single-shot

detectors typically must trade accuracy for the speed increase, as they fail to recognize small objects [69].

Figure 2.3 Object detection results using FasterRCNN. Image taken from HICO-DET dataset.

2.2.3 Human Pose Estimation

One other visual perception task that pertains to this work is human pose estimation. The goal of pose estimation is to locate and identify the different parts of a human body from a static image or video. Pose estimation can be a valuable tool entertainment sports and medical fields, showing similarities and differences in the way's humans move and orient their body parts [11]. For the task of human-object interaction

detection, we are very focused on human body positioning and appendage placement as they can determine the different ways humans interact with objects.

Training of pose estimation algorithms is a highly supervised task requiring large datasets of humans in various activities [19]. The keypoint map acts as the ground truth for evaluation, with points labeled in the image that represent the joints of the human body [18]. These joints include knees, elbows, neck, and ankles. This can be seen in Figure 2.4 with the labeled image of a runner in action. The keypoint map is overlaid over the human with the yellow connecting lines representing the appendages that connect the joints on the runner's body.

There are a few popular approaches to the task of human pose estimation, with almost all based on deep convolutional neural networks. One approach is formulating the problem as a regression problem where joints or key-points are identified and the location prediction error is progressively fed back through the network, as seen in [2], [9] and [54]. Similarly, to this approach the authors of Densepose [1] use a method of semantic segmentation of the human body, to identify the appendages. The second approach to pose estimation is known as the heat-map based approach, where heat-maps are generated to represent pixel-based probability of a key point in that location. The heat map approach can be seen in [11] and [58]. Examples of this method can be seen in [44], [54], and [55]. The output of such models generates a keypoint map as seen in Figure 2.4.

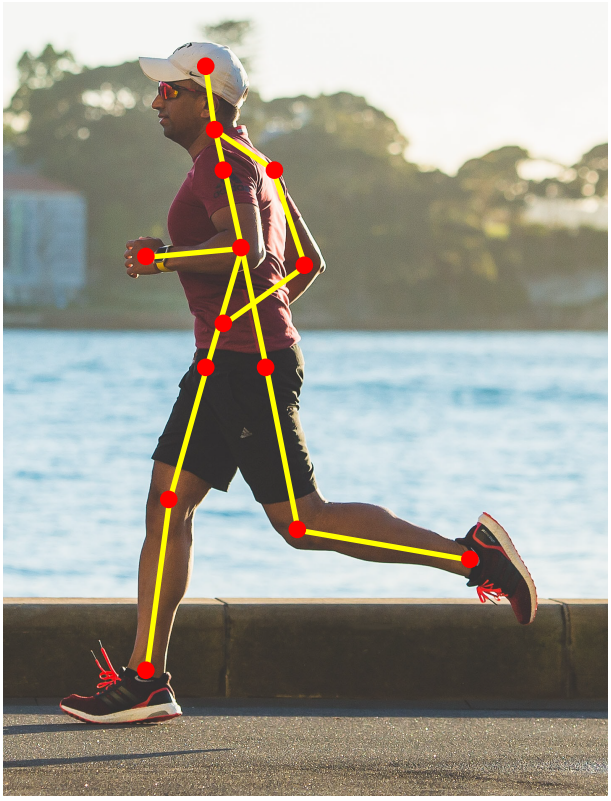

Figure 2.4: Image of a runner overlaid with a human keypoint map, as generated by pose generation algorithms.

2.3 Visual Understanding Tasks

As object detection matured as a research field, interest began to grow in other areas of computer vision. Robust object detectors gave researchers the ability to move beyond perception tasks, and into understanding tasks. In a scene, there could be a lot of information that is not easily identified by simple perception and identification of objects. Visual understanding seeks to extract fine-grained information from a scene. Human-object interaction is an example that fits within the umbrella of visual understanding. Very closely related is the task of visual relationship detection. Human-object interaction detection should be considered a subset of visual relationship detection. A brief overview of visual relationship detection follows and is worth understanding due to the numerous similarities between the two fields.

2.3.1 Visual Relationship Detection

Visual relationship detection seeks to discover the relationships between objects in an image. The discovery of a relationship can commonly be expressed as a triplet in the form of *subject, predicate, object*. While perception tasks like object detection seek a general idea of what is physically present in the image, context reasoning tasks such as visual relationship detection, attempts to find a deeper understanding of what this image means [40]. Take for example the image in Figure 2.4a. Initial perception can identify a man and a horse in the image, but through a better understanding of actions and relationships, we know that this image shows a person riding a horse. In a simple image with a limited number of object proposals, as in the previous example, the number of possible relationships is small. But it is common to have many objects in an image, increasing the size of the search space exponentially. Visual relationships are easily identified by spatial relationships between objects in the image, and many state-of-the-art models heavily weight spatial information between the two objects in the image to attempt to reduce the massive search space that can be present [30]. Another hurdle to visual relationship detection is the long tail distribution that the predicates can exhibit [27] Since most state-of-the-art models are highly supervised approaches, they depend on data previously observed. It is common for datasets to express the real-world commonality of actions or relationships. Given the example images in Figure 2.4b ad 2.4c, human feeding horse is a common image, therefore datasets can have many examples of this relationship. However, consider the relationship human feeding zebra. As a very uncommon relationship, a dataset could contain very few, if not zero of these training examples. However, it does exhibit similarities to the relationship of human

feeding horse. There are many state-of-the-art models and datasets that attempt to tackle the issue of these unseen relationships. The survey by Liu et al. [40], provides a deeper description of visual relationship detection.

Figure 2.5: Examples from the HICO-DET Dataset. a) Human riding horse b) Human feeding horse c) Human feeding zebra

2.4 Related Work

There have been prior works and introductory developments in human object detection such as [12], but we will focus on convolutional neural network based developments. We have classified the methods of solving HOI detection problems into the two classes: multi-stream architectures and graph networks. Multi-stream architectures produce promising results and are easily augmented with supplemental information detection methods such as pose and gaze. This section will provide further insight into how each of these approaches identifies human-object interactions, as well as their strengths and drawbacks.

2.4.1 Multi-stream Approaches

A widely used strategy for creating models that perform well on HOI detection tasks is a multi-stream neural network architecture. Multi-stream convolutional neural networks were first proposed for the task of human object interaction detection by Chao et al. as HORCNN [6]. HORCNN includes three "streams", based around CNN architectures, to extract features from different sources in the image. Using object proposals from the RCNN [16] object detector, the human and object streams extract appearance queues from the image. The human stream can interpret human pose at an elementary level. For example, a person riding a bike is most likely to be in a sitting pose rather than standing. Similarly, the object stream can interpret the appearance of the object involved in the interaction. Again, using the riding-bike example, a bicycle being ridden has a higher probability of being occluded by the person in the image. The final stream in HORCNN extracts spatial information between the human and object. This may be one of the more obvious queues when inferring human object interactions. Reusing the riding-bike example, a human riding a bike is more likely to be located on top of the bike rather than to either side if they were instead standing-next-to-bike. Both the human and object streams are based on CaffeNet [23] implementations, pretrained on ImageNet. Each stream performs a classification for the possible HOIs, and an element-wise sum is taken for their feature vectors for final classification scores. Due to the multi-tasking nature of humans, HOI detection should be considered a multi-label classification problem, as a person can be performing more than one interaction on an object at a time. The individual streams and network architecture of HORCNN can be seen in Figure 2.6.

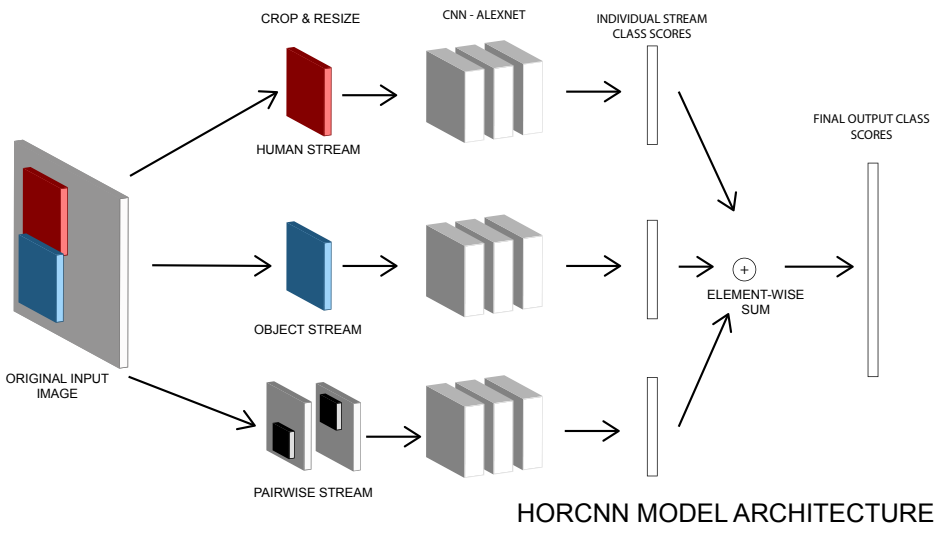

Figure 2.6 Diagram of the HORCNN architecture.

Building on this method of multi-stream approach, Gkioxari et al. [17] use a similar architecture for detecting HOIs, InteractNet. They use three branches based on Faster R-CNN: an object detection branch, a human-centric branch, and an interaction branch. The object detection branch is identical to Faster R-CNN [52]. Bounding box regression for humans and objects is performed as well as computing a classification score for the detected objects. The human-centric branch performs two tasks, action classification, and target object localization. Similarly, to HORCNN, human appearance is used to compute an action classification score or the probability that the human in question is performing a specific action. Target localization again uses human appearance features to the probability density of the action's target object location in the image. The final branch of interaction recognition combines the features detected for the human-centric branch with appearance features from the target object. The score is computed by preforming sigmoid activation from the outputs from human action and target action classification which are represented as vectors. For inference, InteractNet uses a cascaded

inference strategy, in which rather than computing the scores for every single action-object pair, they compute the bounding box for the object that maximizes the score for a specific action. For actions with no object interaction, the score from action classification in the human-centric branch is used. A diagram of the InteractNet architecture can be found in Figure 2.7.

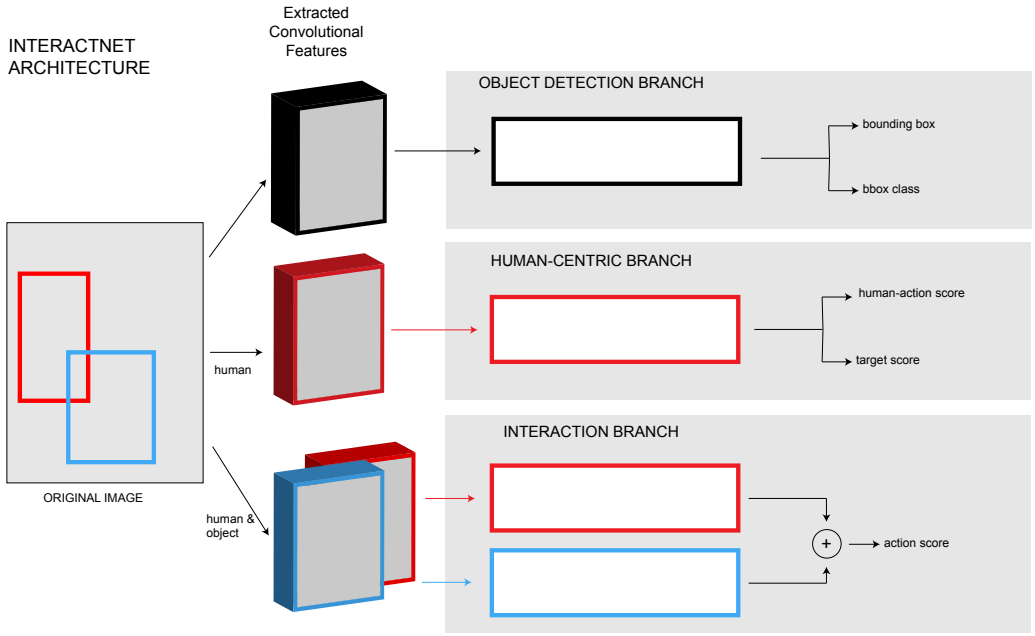

Figure 2.7 Architecture from InteractNet.

Another implementation of the multi-stream architecture is presented by Gao et al. instance centric attention network (iCAN) [14], proposes using an attention-based mechanism for their architecture streams. As seen in HORCNN, the three streams used are a human, object, and spatial configuration stream, and generating proposals from the Faster R-CNN detector. The difference in the streams from HORCNN is the use of the proposed instance-centric attention network, replacing the conventional CNN architectures. Unlike extracting object appearance and human appearance as individual

queues, iCAN aims to extract contextual features from both the human and object instances in the image. iCAN begins by extracting the appearance features from the localized object to dynamically generate an attention map on that object instance. This is accomplished by embedding the appearance features and convolutional feature maps and measuring similarity using a dot product operation. The attention map is generated using a softmax function. A contextual feature is extracted from the attention map through the weighted average of convolutional features. The iCAN module outputs a concatenation of the instance level appearance features and the contextual appearance features. A diagram of the architecture for the iCAN module can be seen in Figure 2.8. Scores for each action are computed similarly to InteractNet and treated as a multi-label classification problem.

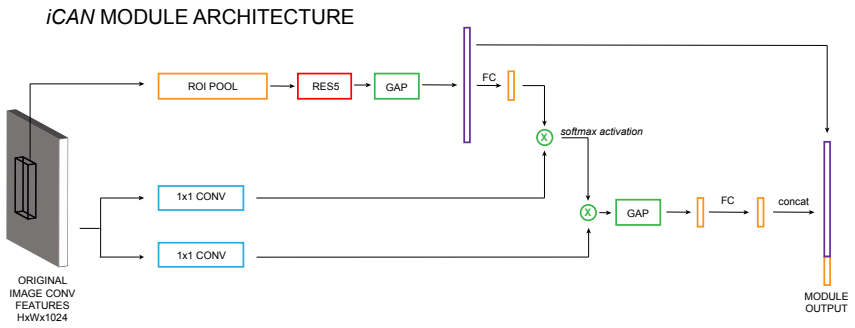

Figure 2.8 Architecture of the iCAN module. GAP signifies global average pooling, and res5 signifies the 5th residual block.

2.4.2 Fine-Grained Information Retrieval

It can be seen from the iCAN implementation that more information than appearance and spatial relations benefit the goal of HOI detection. There has been

considerable research into using finer- grained contextual information extracted from the detected human to enhance HOI models. Researchers have experimented with language models providing supplemental information such as in [3], but we will focus on visual information extraction. Pose information from the human in the image can supply very important characteristics specific to that action [36]. One of the models to investigate these methods is a model using the individual body-part attention as proposed by Fang et al. [13]. The authors note that just using individual body-part attention does not capture the correlation between different body parts used in a specific interaction. Therefore, they propose generating attention maps from pairs of body parts and select specific pairs that best fit the interaction in question. Many works have proposed using human pose estimation to aid in detection results, some of the first being Gupta et al. [21] and Li et al. [34]. Li et al. propose generating a heat map of human joint key-points in their model referred to as the Interactiveness Network, which is used as an add-on module for existing HOI detection models. This module uses three streams as in HORCNN, with appearance features from humans and objects. The difference is in the spatial information stream, where the pose map is incorporated with the spatial configuration map. A convolutional architecture is used to extract the feature representing both pose and spatial configurations. This output is concatenated with the human and object streams to create an interactiveness score, which is integrated with the interaction classification scores from an existing model. It should be noted that the interactiveness score only applies to HOIs in which the human physically interacts with an object to produce the interaction. Therefore, only these interactions can benefit from this method. Li et al. also incorporate a knowledge transfer training mechanism that influences the Interactivness Network

module. This mechanism provides learned information from multiple datasets to produce a highly accurate inference on a testing image.

Another model that uses pose estimation is the Pose-aware Multi-level Feature Network or PMFNet proposed by Wan et al. [56]. This approach utilizes a different architecture than previously examined in this survey. PMFNet builds upon the method of body part attention maps, but not constrained to pairs as in the Interactiveness Network. Additionally, spatial relations between body parts and the object in question are computed to encode fine spatial configuration information. The multi-stream architecture employs three modules, a holistic module, a zoom-in module, and a fusion module. Using human, object, and union (interaction area) proposals detected using Faster R-CNN [52] as an object detector, a conventional CNN architecture is used to extract appearance features. This same CNN also extracts a spatial configuration map between the human and the objects. The authors use the CPN pose estimator developed by Chen et al. [9].

The spatial features, appearance features, and pose estimation are fed to the holistic and zoom-in modules. The holistic module aims to capture object level and related context information. It consists of four streams: human, object, union, and spatial con- figuration. Each stream is responsible for embedding respective output features. These are concatenated to create a holistic feature representation. The zoom-in module is responsible for extracting fine-grained information from the human pose spatial configuration. These are considered human body part-level features. This module contains three branches that extract human part level appearance features, human part level spatial configuration features, and an attention component to enhance relevant human parts to each specific interaction. These features are concatenated to result in the

local feature representation. In the final fusion module, both the local features and the holistic features are used to fuse relation reasoning from both the coarse level and fine level features. The first benefit of this module is the ability to use coarse features as a contextual cue to suppress interactions that cannot exist in the current set of human and object proposals, this is denoted as an interaction affinity score. The other benefit is an ability to use both object level and part-level features to determine the relation score from fine-grained representations, denoted as the local relation score. Both the interaction affinity score and the local relation scores are fused to create a final score for the interaction given the human and object proposals. PMFNet is trained in an end to end manner using cross-entropy loss, with the exception of the Faster R-CNN [52] and the CPN [9] modules.

One method of note proposed by Xu et al. [60], intention driven human object interaction detection or iHOI, incorporates the features obtained from human gaze following. This is done through another multi-stream architecture. First, a set of visual and spatial features are extracted using established methods. As is common in human-object interaction detection, Faster-RCNN [52] is used to create human and object proposals. A pose estimation network from [9], and a gaze direction detector borrowed from [55], are trained on other datasets and used to extract human body joint locations and gaze target location respectively. These features are combined into three separate streams in the model. An individual stream for extracting appearance features from both the human and object, a human-object pairwise stream for extracting features from the spatial configurations and appearances of the human and the object together, and finally a gaze driven context-aware branch that aims to infer the focus area of the human through

body positioning and through the gaze location. These features are then combined to create a final human object interaction prediction. However, iHOI does not improve performance of human-object interaction detection much beyond its contemporary counter parts. There has been some discussion of integrating more modern gaze following algorithms such as [45], [62], or [64] However, these approaches are considered slow, needing many network streams and extra processing to make a final prediction.

A recent model by Zhou et al. in [67] proposes a very complex multi-stream network architecture, incorporating language priors, geometric features, and visual features to achieve a high score on the V-COCO dataset. Their visual feature module includes using gaze type cues as well as pose estimation features to create a very robust prediction based on just the visual information present in the image. The geometric feature branch is strikingly similar to the spatial or pairwise streams of previous models like [6] and [17]. Another work called Parallel Point Detection and Matching (PPDM) [37], use purely spatial features to predict the interaction class between humans and the objects. They also implement a novel hourglass shaped neural network backbone for their model. PPDM performs well on HICO-DET dataset.

2.4.3 Graph Neural Networks

An image with human object interactions can be interpreted similarly to a scene graph [24], in which the nodes represent objects and humans while the edges connecting the nodes represent relations between them. A comprehensive survey of graph neural networks and their use in visual understanding tasks can be seen in the work by Wu et. al [59]. This method is very similar to the task of scene graph generation, such as the work

seen in [24], which is followed very closely in the human object interaction detection task by [61], breaking the task down into a graph. Qi et al. [50] propose a novel model using a graph neural network based on message passing. The goal of the model, called the Graph Parsing Neural Network (GPNN), is to take a complete human-object graph of the image that includes all possible interactions between the human and the objects and remove edges that represent non-existing interactions in the image. This structure enables the model to preserve spatial relationships while detecting human-object interactions. GPNN generates the graph structure through the use of a link function. Then the message, update, and readout functions are used in belief propagation. The message function is used to summarize messages or information coming from other connected nodes, while the update function updates the hidden node states according to the incoming information. The final readout function generates an output label based on the hidden node states. Each function uses various neural network architectures as detailed in their paper. The probability of an HOI occurring between nodes is a product of the final output probabilities between the human and object nodes.

Using the idea of graph neural networks, Zhou et al. [66] provide an improvement on the GPNN [50] model. Known as the relation parsing neural network (RPNN), this network focuses around two graphs, an object body part graph and a human body part graph. The object body part graph describes the relationships expressed in the image between body parts of a specific human and the surrounding objects in the image. The human body part graph models the relationship between the human and their body parts, similar to the task of pose estimation, to describe the actions and movements of the human as they relate to a specific interaction. The two graphs are fused using a message

passing mechanism like in GPNN to convey information for a final interaction class prediction. This network body part contexts to predict actions. RPNN performs very well on HICO-DET and V-COCO. A more recent work into graph neural networks was conducted by Liang et. al [35], earning this paper a top mAP score for the HICO-DET dataset. However, unlike GPNN, they use a dual graph strategy with semantic information coming from the class labels and visual information to construct a final optimized scene graph of each object and human in the image. This model currently has the highest performance score on the HICO-DET dataset. Graph neural networks seem to be outperforming other methods for human object interaction detection, there have been many recent works that exploit them as well as other information such as pose estimation, [65] is a good example of this.

2.4.4 Weakly Supervised and Zero-Shot Approaches

An interesting area of computer vision research is in the area of weakly supervised and zero-shot approaches to learning. Weak supervision entails that a learning algorithm is given very few training examples of a specific task, such as identifying objects. Zero-shot signifies that the specific example has never been seen by the algorithm. Both weak supervision and zero-shot approaches for more classical tasks of computer vision, such as object detection [33], have been well documented throughout the years, even without the use of deep convolutional neural networks as in [5], and [7], and using autoencoders as seen in [29]. Interestingly [28] uses information learned from the task of human object interaction detection to aid in the task of object detection.

Specifically, for human object interaction detection, zero-shot and weakly supervised learning techniques are useful due to most datasets expressing a long-tailed

distribution of image data. The long-tailed distribution describes the greater prevalence of common examples in the data than that of more uncommon examples. For example, there are many more examples of human-ride-horse than examples of human-ride-zebra, both because of the rarity of zebras and the rarity of scenarios where a human would be riding a zebra.

However, the example of human-ride-zebra is not an impossible scenario, and a well generalized model should be able to identify these rare relationships just as humans can. This long-tailed distribution in datasets reflect the real-world, where we know that some interactions are rarer than other. For visual understanding tasks this process becomes more difficult as it is harder to rely on well-defined visual features such as those generated by SIFT [42] features or convolutional neural networks [63]. However, some distribution issues can be attributed to the dataset, as seen in the study [27], exploring HICO-DET and some of the multi-stream models covered in this survey. An attempt at the task of zero-shot recognition and weakly supervised learning is seen by Pyere et al. in [47], incorporating semantic language information from large text databases that provide probabilities for the interaction in question. One very early example of a weakly supervised approach is seen by Prest et al. in [49] using a probabilistic type model, however it has not been tested on modern datasets such as HICO-DET.

More recent work seems to focus on improving these zero-shot interaction classes, and these improvements even help overall generalization on most datasets, this improvement can be seen in works such as [4], [25], and [57]. Hou et al. [22] propose the visual compositional learning framework for human object interaction detection. Their network learns shared object and verb features, breaking down verbs to relate to specific

objects. This process learns shared object and verb features from across all human object interactions. Their framework uses another multi-stream process containing three streams. Specifically, their main contribution is their verb- object branch that extracts verb or interaction class features from the union of both the human and object bounding boxes. They show superior performance on the HICO-DET dataset using this method. Another interesting recent work on improving generalization across the lesser seen interaction examples is done by Song et al. in [53]. They propose using adversarial domain generalization to encourage predictions on the unseen or longer tailed examples. Specifically, they focus on improving the spatial stream in a network similar to that of HORCNN [6] as this branch is object invariant by design. They create a type of zero-shot learning dataset by reorganizing examples in the training and test sets of HICO-DET [6] and using parts of the UnRel dataset [48] as a validation set. They do show great performance on zero-shot interaction categories, however we cannot rank their approach as they do not rank their improvements against other models on HICO-DET. They propose their learning framework as an add-on to existing models.

We show the mAP scores on the HICO-DET dataset for most of the key models covered in this section in Table 2.1. The scores listed were found by their authors and listed in their papers. We can see the performance improvement by adding finer-grained features from the image to the prediction models. HICO-DET offers several evaluation setups and difficulties shown in this Table. More information on this can be found in section 2.5.4.

Table 2.1: Model comparison evaluated in %mAP

	Default Setting			**Known-Object Setting**		
Model	**Full**	**Rare**	**Non-Rare**	**Full**	**Rare**	**Non-Rare**
HO-RCNN [6]	7.81	5.37	8.54	10.41	8.94	10.85
InteractNet [17]	9.94	7.16	10.77	-	-	-
GPNN [50]	13.11	9.34	14.23	-	-	-
iCAN [14]	14.84	10.45	16.15	16.26	11.33	17.73
Interactiveness Net [34]	17.03	13.42	18.11	19.17	15.51	20.26
PMFNet [56]	17.46	15.65	18.00	20.34	17.47	21.20
VS-GAT [35]	20.27	16.03	21.54			

2.5 Datasets and Evaluation Metrics

This section introduces the most common datasets used in the task of human object interaction detection and provides insights on how they differ. Machine learning models rely on previously seen data to guide predictions for a specific inference task. Therefore, the quality and quantity of the data the model learns from are important for making good predictions. High-quality datasets commonly contain localization and class labels on each of the objects or humans in the image. Human object interaction detection requires image data to be labeled not only for objects but also for the relationships between the human and objects. For images with many instances of an interaction, these all must be separately labeled. Human object interaction datasets must contain enough training data for all object classes as well as all relationship classes. Data for all possible real-world combinations of objects and relationships are impossible to obtain, therefore datasets typically pick a number of objects and interactions to focus on. There are many

datasets used for this task, however, each dataset uses specific methods of providing ground truths, as well as different object and interaction classes. Each dataset also provides its own method of evaluating model performance. Table 2.2 summarizes the datasets and their properties, as discussed in this section.

Table 2.2: Summary of dataset properties

Name	Images	Interaction Classes	Object Classes
HICO	47,774	600	80
HICO-DET	47,776	600	80
V-COCO	10,346	26	80
HCVRD	52,855	927	1824

Arguably, one of the first purpose-built datasets for the task of human object interaction detection is the HICO [7] dataset, created by Chao et al. This dataset was constructed from the MS-COCO [38] dataset commonly used for object detection evaluation. HICO uses 80 object categories from MS-COCO and commonly used verbs to create the interaction categories for each object. Each object is also given a "no interaction" action, for a total of 600 human-object interactions. Each human object interaction category has at a minimum of six images, and the test set should contain at least one image for that category.

The HICO dataset does not provide instance level groundtruth annotations for every HOI occurring in each image. Another problem is the fact that images with multiple humans present are not exhaustively labeled. For example, in the case of a

person-riding-in-airplane, there could be many people seated on board an airplane in the image, yet the HICO dataset would only require detecting a single HOI that fits that description. That is to say, that the HICO dataset proves image level groundtruth annotations. With these issues in mind Chen et al., the same authors of the HICO dataset, augment HICO to create HICO-DET [6]. HICO-DET contains groundtruth labels for every human, and object participating in an annotated interaction class. The authors took the original HICO dataset and augmented it by crowd-sourcing the instance level groundtruth labeling via Amazon Mechanical Turk.

The verbs in COCO (V-COCO) dataset [20], is another commonly evaluated dataset for human object interaction detection. Similar to HICO, the object classes are taken from the COCO [38] dataset. But unlike HICO, the authors use the images already found in the COCO dataset. COCO has human-labeled and verified captions on each image, these are where the interaction classes are derived from. Using a simplified vocabulary, they designate 26 common actions amongst the different object classes. The COCO dataset contains ground truth labels for each object and human in the image, and the authors of V-COCO were able to reuse these.

Another dataset, although less commonly used, for human object interaction detection is the HCVRD dataset created by Zhuang et al. [68]. This dataset is far more diverse in terms of labeled interactions and objects than the previously covered datasets. The images for HCVRD were gathered from the Visual Genome dataset [31], which contains object labels and bounding boxes, image captions, and labeled relationships between objects. The interactions included in HCVRD were drawn from the VG dataset where one of the objects is labeled as human. The authors took special care in "cleaning"

the interactions by removing ambiguous actions and combining interactions with close similarity as a single interaction class.

In human-object interaction detection, mean average precision (mAP) is most commonly used as an evaluation metric. For each image, the model should output a classification score for each interaction class. For each class, average precision is calculated from the entire test set of images. The mAP is computed as the average of the average precision scores. The authors provide an easy setting for evaluation called the "Known Object" setting. In this setting the verified positive images are used as positives with the verified negative images used as the negatives, skipping both the unknown and ambiguous images [7]. This removes the uncertainty of an imperfect object detector, by removing the images without the subject from the human object interaction in question. For a more realistic setting, the authors propose adding the unknown category of images back as extra negatives. Benchmarking on the HICO and HICO-DET datasets are done on both the Known Object setting as well as the realistic setting.

Two common metrics for evaluation of models on the V-COCO dataset are agent detection and role detection [20]. For agent detection, the task is to detect the humans performing a queried action. Average precision is used in this task as a performance metric, where humans labeled with the correct interaction category are marked positives. For role detection, the goal is to detect the human and objects participating in the given interaction. A model should produce a bounding box for the human and for the role. Using the intersection-over-union between the detected bounding box and the ground truth labels, average precision is computed and scored as the metric for this task. Models trained on HCVRD are tested against three metrics: predicate recognition where the

interaction is detected given the bounding boxes for the human and object. Phrase detection in which, given the human and object bounding boxes, the interaction as well as a union bounding box that encompasses the entire interaction or activity is predicted. For the final test metric relationship detection, measured in terms of recall, the model must localize the human and objects, as well as perform phrase detection.

One last dataset to mention is the UnRel dataset [48]. UnRel is specifically created to evaluate unrealistic relationships between objects and people. However, it specifically focuses on spatial relationships such as person-ride-dog or elephant-on-top-of-car and includes non-human object interactions. It can be used for add-on module training or in the case of [53] where they manually filter out interaction classes that do not pertain to humans, as supplemental data. It is worth mentioning that a dataset of unrealistic interactions could help benefit future zero-shot and weakly supervised learning approaches to human object interaction detection.

CHAPTER III

METHODS

3.1 Building the Toolkit

We use the PyTorch deep learning framework [46] as it is standard in industry and academia for machine learning research. The primary motivation behind PyTorch is the ease of use compared with other deep learning frameworks such as Caffe, MXNet, and Tensorflow. All of these packages are centered around the idea of automatic differentiation to efficiently compute the gradients in the model for the essential gradient based optimization at the core of deep learning. A simple yet efficient Python API for this framework makes creating and setting up training routines with PyTorch relatively easy and straightforward, freeing up researchers from software engineering tasks and allowing them to focus on advancing research in their fields. In order to avoid trading speed for ease of use, the core of the PyTorch framework is written in C++ for fast execution compared to native Python programs. It also enables the use of hardware acceleration from GPU (graphics processing unit), utilizing the massively parallel architecture to provide the computational power for training deep learning models, all without the need for complex graphics programming constructs such as those supplied in Nvidia's CUDA packages. PyTorch also includes streamlined multi-GPU integration into their API, for systems utilizing more than one hardware accelerator. Since the opensource release of PyTorch in 2017, the number of deep learning researchers that to transferred to this framework is increasing exponentially, making it the top choice for deep learning frameworks. One of our goals for this project is to provide researchers with an easy to use

set of tools to advance the research in the area of human-object interaction detection and using PyTorch as our deep learning framework seems like an obvious choice.

We chose to implement a data loader for the HICO-DET dataset using PyTorch. HICO-DET is commonly used as a benchmarking dataset for human-object interaction [43] detection, and its related precursor HICO was a foundational dataset to the field. This dataset is publicly available online, utilizing MATLAB for performance analysis. To integrate HICO-DET into our toolkit, we began by converting the annotations in the dataset from the supplied MATLAB files to data structures that are easily parsed and read by Python. However, the annotations supplied by the dataset are fairly complex. Bounding boxes for humans and objects engaging in a specific interaction in the image are listed, and multiple humans and objects can participate in an instance of this interaction. Therefore, the connection is utilized to connect two bounding boxes as participating in an interaction. We treat each one of these connections as a separate proposal.

After assembling the ground truth annotations, while training and testing, an object detector needs to search the image being sampled for humans and objects. As mentioned, HICO-DET uses the same classes of objects as found in the popular MS-COCO [38] dataset. This allows us to use any object detector pre-trained on MS-COCO, as the head of our data pipeline. We integrate an option for the user to add their chosen object detector as the proposal generator, or supply annotations any other way. The bounding boxes proposed by the detector are passed to the data loader to be used as proposals. We set up our toolkit training for optimal training speed and use of the GPU memory. Keeping the object detector and the interaction detection model in GPU

memory during training could be infeasible due to the amount of resources a specific system has available and would slow down training considerably. Since we are not training the object detector, we precompute a list of human and object proposals to keep ready when those instances are needed during training. This should speed up training times considerably.

3.2 HORCNN Implementation

For the HORCNN model, and various others surveyed in chapter 2, an image-centric sampling strategy is used, where each batch contains a fixed number of human-object proposals from a single image [6]. Our data loader follows a similar strategy, enabling the user to specify a specific number of proposals per image to use in a batch. Furthermore, the image-centric sampling strategy uses proposals of three types; true positives where both the human and object proposal box have an Intersection over Union (IoU) overlap greater than 0.5 and the interaction label contains the object in the box, type-1 negatives where the IoU is between 0.1 and 0.5 again where the interaction in question contains the object in the bounding box and type-2 negatives where the object in question is not included in the bounding box. A random distribution from the three proposal classes form the image proposal batch. Furthermore, PyTorch allows the user to easily specify the batch size for training the model, which in this case, would be the number of images to sample from. Our HICO-DET PyTorch data loader allows for us to load the training and testing set in the same manner, given user specified constraints on the batch sizes.

The task of human-object interaction detection should be considered a multilabel classification problem, as humans can perform multiple actions on an object at once. To

generate ground truth vectors for each positive prediction, we must search the dataset for similar bounding boxes, identified by an IoU > 0.5 for both human and object, and provide a positive label for the interaction class in the ground truth. We create these labels before training the human-object interaction detection model before training to speed up the process.

To demonstrate the dataset, and to add a baseline model to the tool kit, we implemented the HORCNN detection model in PyTorch as a baseline comparison model. Implementing in PyTorch allows researchers to test their own models against a baseline implemented in the same framework, as long as they use PyTorch. Following the implementation details, we re-created the model, which was originally built using Caffe. Other than the change in framework, a few deviations should be noted. The original authors use Fast-RCNN [52] or individual RCNN detectors for each object for their object detectors. While these are fairly accurate and robust detectors, Faster-RCNN has been proven to outperform both. Using a high performing object detector is important for this task since the very first step in the detection pipeline is object detection. As previously stated, the data loader allows the user to integrate any object detector they choose so that improved algorithms can be used in the future. We chose to use Faster-RCNN due to its immediate availability as a module included with PyTorch, pre-trained on the MS-COCO dataset which uses the same object classes from HICO-DET. Another deviation is the use of AlexNet [32] rather than CaffeNet [23]. For their implementation of HORCNN, the authors use CaffeNet pre-trained on the ImageNet dataset for the human and object convolutional streams. For all intents and purposes, AlexNet and CaffeNet are the same architecture, with CaffeNet being modified for single GPU use. To

avoid having to perform costly ImageNet pretraining, we used the pre-trained AlexNet implementation provided by Torchvision, widening the output feature vector to 600 classes to match the output classes of the HICO-DET dataset. We will make available several versions of pre-trained HORCNN models for researchers to perform baseline tests.

CHAPTER IV

RESULTS AND DISCUSSION

4.1 Performance of Model

For the re-implementation of the HORCNN model, we see a close but slightly reduced mAP on the HICO-DET dataset, close to that of the original papers. Differences could be explained by hyperparameter adjustments. Due to computational constraints, our implementation was trained with a batch size of four images, containing four randomly sampled proposals from the true positive, type I negative, and type 2 negative proposal sets, listed in the previous section of this paper. Using the batch sizes results in a total batch size of 16 proposals. In the original work, eight images are selected per batch, with 8 proposals per image, for an overall batch size of 64 proposals. We trained four times as long as the original work due to the reduction in proposals from our training parameters. We trained for 400k iterations at a learning rate of 0.001, and 200k iterations at a learning rate of 0.0001. Results and comparisons can be seen in Table 4.1. The model was trained for ~20 hours on a single Nvidia TitanXp GPU.

Table 4.1: Performance (%mAP) of re-implemented model vs. published results

Model	**Full**	**Rare**	**Non-Rare**
Ours	5.87	3.06	7.08
HORCNN	7.81	5.37	8.54

4.2 Performance of Individual streams

The three streams of the HORCNN model, human, object, and pairwise streams, extract fine-grained features from their subjects. However, these feature weights are

summed when making a final prediction on whether a human-object pair is engaged in an interaction. From a general understanding of human interaction, we know that fine-grained features such as body placement and pose can influence a decision on whether or not a human is interacting with an object. We perform studies on each individual stream and selected combinations to see if one performs best in the overall task of identifying an interaction. We evaluate performance similar to the authors of HORCNN, by following the mAP criteria from the PASCAL VOC classification competition. Results of these test can be seen in Table 4.2.

Table 4.2: Performance of the individual model streams (%mAP)

Model	**Full**	**Rare**	**Non-Rare**
Full Model	5.87	3.06	7.08
H	1.62	0.40	2.09
O	4.65	2.78	5.67
P	0.93	0.07	1.08
HO	5.41	3.51	6.54
HP	1.41	0.15	1.65

4.3 Dataset

The HICO-DET dataset is large and fairly diverse, however, there are a few issues present. First, for each object category, there is a ‘no interaction’ class. This provides samples for a model to learn how to distinguish when there are objects and humans in an image, but they are not interacting with each other. However, there are many more instances where there should be a no interaction category, but they are not labeled. Many of these instances stem from objects or humans that are detected by an object detector,

such as Faster-RCNN used in this study, but are not present in any ground truth as participating in an interaction with a human. While training with the image-centric sampling strategy, the model could be given these samples, since samples are chosen at random, without a label and will be penalized in the loss function since no ground truth exists. It is possible to hand label these human-object proposals with a 'no interaction' proposal while loading the data, but in doing so the dataset becomes imbalanced. Interestingly, there are some human-object pairs that are participating in an interaction class in images in the dataset, that are not labeled. For example, the image in Figure 4.1 is taken from the HICO-DET training set with bounding boxes representing detections from Fast-RCNN. The ground truth annotations only contain labels for four separate humans 'sit at' and 'eat at' dining table. But clearly, we see that one human is drinking from and holding a cup, as well as many cups and plates in the image that should be labeled with 'no interaction'.

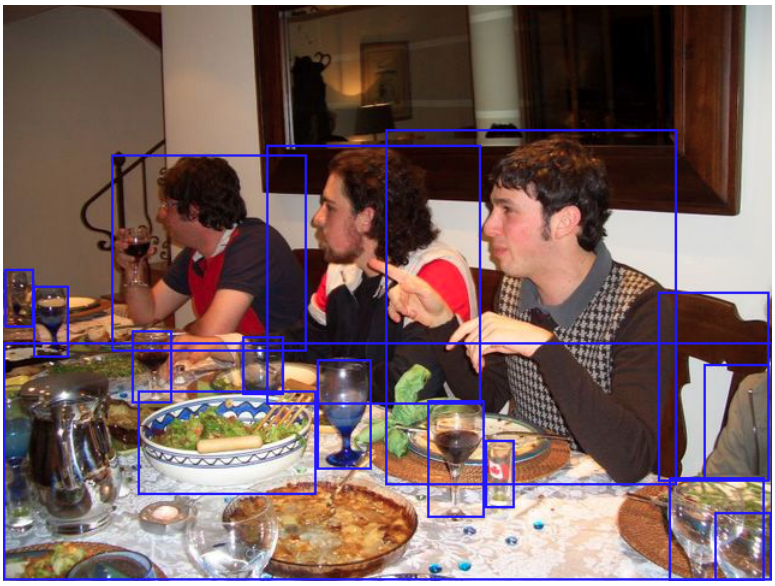

Figure 4.1: Example of non-exhaustively labeled image from HICO-DET

The fact that human to human relationships are present in the HICO-DET dataset, provides an extra complexity when searching for proposals. Unfortunately, all images containing multiple humans does not have a 'no interaction' label between these human detections. Since the object detection selection must pair all humans with all objects, and all humans with all humans, it is likely that one of these unlabeled human to human relationships show up in the dataset. While it is possible to create these labels artificially in the data loader, it adds more unnecessary data preprocessing for the training. And it is not guaranteed that these labels are true 'no interaction' labels, instead of missed interactions. Figure 4.1 has examples of these missed interaction labels. We see multiple humans present, but as noted, the ground truths only contain human-dining table interactions.

HICO-DET contains a number of rare human-object interaction classes, as evidenced by the 'rare' setting for evaluation. However, the quality of these examples leaves doubt in the ability of the human reviewers to filter out poor images, or images that do not display the interaction. For example, the image seen in Figure 4.2 contains training and test images labeled as containing the relationship of human-repair-mouse, mouse in this context referring to a computer mouse. It is clear from this picture that there is no human present in the image. An automated data-processing pipeline would not label this as the interaction class human-repair-mouse, and more troublingly, this is the one of two training examples for this interaction in the entire dataset. This issue could be present in other small objects in the dataset; however, we find this to be the most egregious error. This brings into question the quality of the HICO-DET dataset, and its ability to train high performing models for human object interaction detection.

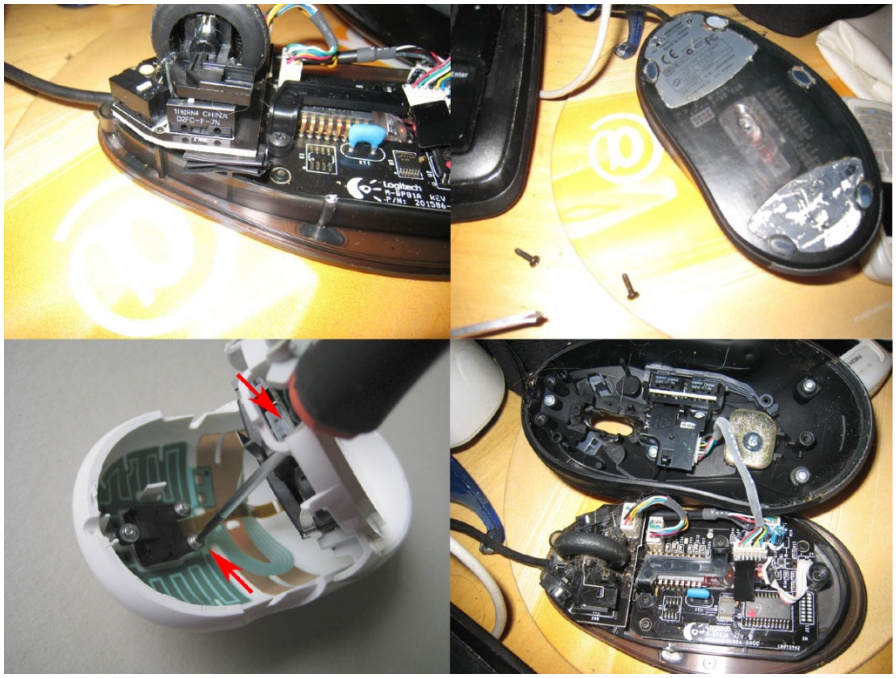

Figure 4.2: Examples of the interaction class *'human repair mouse'* from the HICO-DET dataset.

Lastly, the image annotations for HICO-DET contains lists of interaction classes present in each image. Each interaction class present will have a list of humans and a list of objects participating in it. In the common case where there exists a human-object pair where two or more interactions exist, the annotations seem to have been performed separately for each interaction, shown by slight differences between bounding box coordinates. Figure 4.3 shows an example of this, where both 'bboxhuman' annotations refer to the same image, but the bounding box dimensions are off just slightly. This bounding box misalignment leads to greater data storage costs, and data-preprocessing requiring IoU (intersection over union) computation on each detection proposal and all other proposals. We suggest a better annotation scheme for the HICO-DET dataset as seen in Figure 4.3. Keeping a list of each human and object bounding box separate from

the interaction class label could make processing more straightforward and reduce training errors.

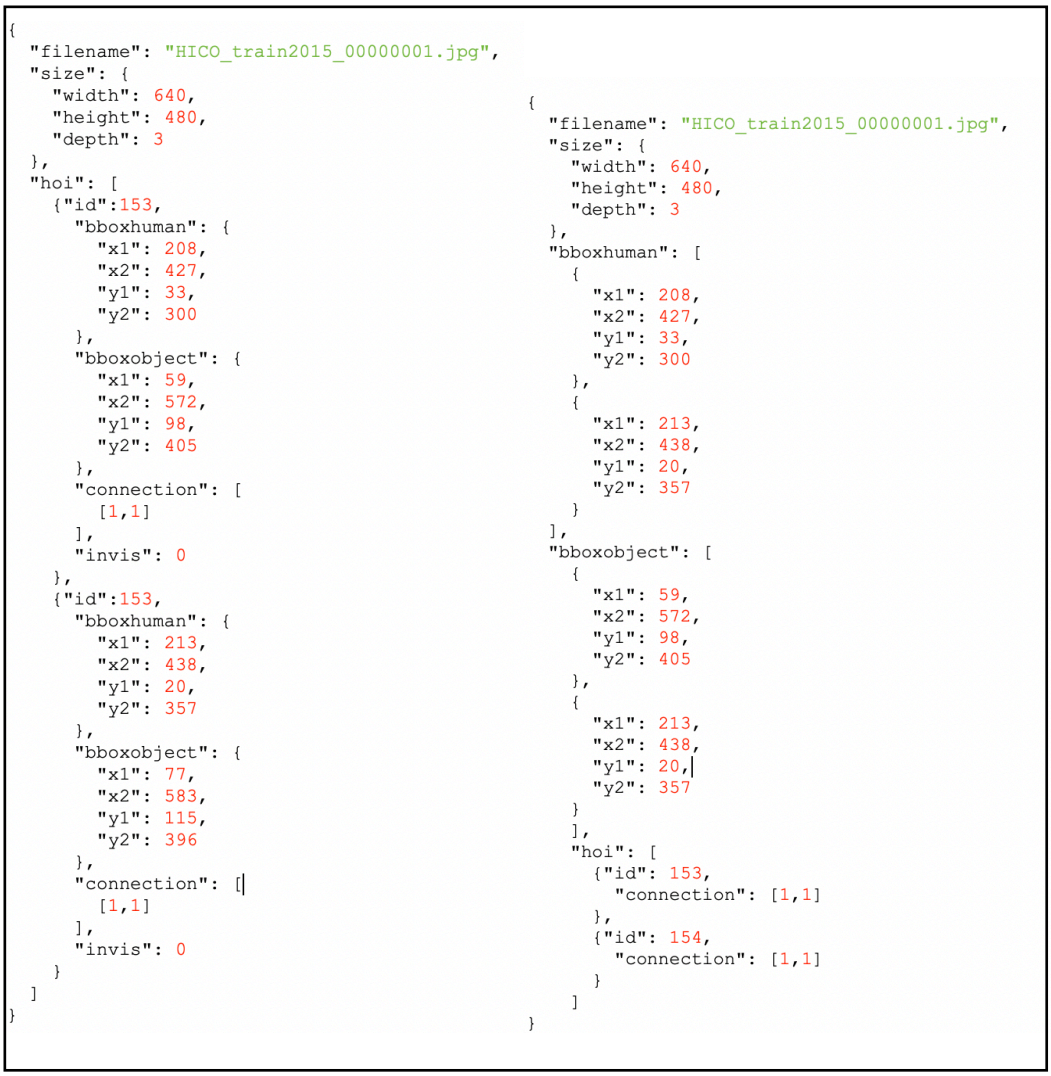

Figure 5.1 Data set annotations, JSON format. a) Left, current HICO-DET annotation scheme, b) Right, proposed annotation scheme

4.4 Model evaluation

We evaluated several test cases for mAP score. The evaluation was done over the entire testing set, using 10 proposals from each image, similar to how the authors of HORCNN perform their evaluations. These cases and their results can be seen in Table

4.2. HOP denotes the full model including scores from the human, object, and pairwise streams. H, O, and P denote human, object, and pairwise streams respectively. HO denotes the score of the human and object branches combined. Finally, HP denotes the human and pairwise streams combined.

Our original hypothesis was that the human stream would be more dominant in guiding predictions, however, the results show that the object stream has the best mAP on the test set and seems to be the dominant factor in the HORCNN model. We believe that this is caused by similar interactions between multiple object categories. For example, the interaction 'carry' is valid for 32 of the 80 object categories. While many of the human appearances could be similar for certain groupings of objects, it is likely that there is not enough information from the human appearances alone to differentiate between these exact object interaction classes. This can be seen in some of the results from test images on the trained model, where similar interactions between objects receive relatively high scores. When combining the human and the object streams, we see that the mAP improves slightly over just the object stream, however, it performs better against the combination of the human and pairwise streams. This shows the importance of the object stream in making predictions on the HOI classes. Unsurprisingly, the full model incorporating all the streams achieves the highest mAP score, this proves the importance of incorporating all three streams in the HORCNN model. Out of the previous works surveyed in the related works section of this paper, HORCNN achieves the lowest mAP scores, quite low for a good prediction model.

CHAPTER VI

CONCLUSION

In this work we have taken an in-depth examination of the task of human-object interaction detection, covering datasets and the baseline models. We performed studies on the baseline model for the HICO-DET dataset, HORCNN, to identify the most robust model components and features. We see that for the multi stream approach presented in HORCNN, the object appearance features provide the most accurate prediction on the dataset. However, it does not compare to the combination of the streams to provide accurate human-object interaction detections. We hope that the findings of these studies can influence future model design in this field of research. The HICO-DET dataset for human object interaction detection was also examined throughout this work. We have shown some concerning quality issues regarding this dataset. It is our opinion that this dataset should be more carefully examined for accurate labeling and higher quality images, especially for the crucial training segment of the dataset. With some updating, this dataset could become very valuable to researchers in this field.

We have presented a basic set of tools for evaluating human-object interaction detection models using the PyTorch framework. We hope that these tools can be used by future researchers to evaluate their models against the baselines presented in the HICO-DET paper. Although out of the scope of this current work, we would like to extend this toolkit to include more models to compare against. Providing these models would give researchers more baselines to compare against and examine. We would also like to include more human-object interaction detection datasets for easier evaluation. We hope that these contributions will accelerate the growth in this field.

REFERENCES CITED

[1] R. Alp Guler, N. Neverova, and I. Kokkinos, “Densepose: Dense human pose estimation in the wild,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7297–7306.

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, “2d human pose estimation: New benchmark and state of the art analysis,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3686–3693.

[3] A. Bansal, S. S. Rambhatla, A. Shrivastava, and R. Chellappa, “Detecting human-object interactions via functional generalization.” In AAAI, 2020, pp. 10 460–10 469.

[4] A. Bansal, K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran, “Zero-shot object detection,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 384–400.

[5] H. Bilen, M. Pedersoli, and T. Tuytelaars, “Weakly supervised object detection with convex clustering,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1081–1089.

[6] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, “Learning to detect human-object interactions,” in 2018 IEEE winter conference on applications of computer vision (wacv). IEEE, 2018, pp. 381–389

[7] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, “Hico: A benchmark for recognizing human-object interactions in images,” in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1017–1025.

[8] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, et al., “Mmdetection: Open mmlab detection toolbox and benchmark,” arXiv preprint arXiv:1906.07155, 2019.

[9] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, “Cascaded pyramid network for multi-person pose estimation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7103–7112.

[10] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, “Deformable convolutional networks,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 764–773.

[11] Q. Dang, J. Yin, B. Wang, and W. Zheng, “Deep learning based 2d human pose estimation: A survey,” Tsinghua Science and Technology, vol. 24, no. 6, pp. 663–676, 2019.

[12] V. Delaitre, J. Sivic, and I. Laptev, “Learning person-object interactions for action recognition in still images,” in Advances in neural information processing systems, 2011, pp. 1503–1511.

[13] H.-S. Fang, J. Cao, Y.-W. Tai, and C. Lu, “Pairwise body-part attention for recognizing human-object interactions,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 51–67.

[14] C. Gao, Y. Zou, and J.-B. Huang, “ican: Instance-centric attention network for human-object interaction detection,” arXiv preprint arXiv:1808.10437, 2018.

[15] R. Girshick, “Fast r-cnn,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.

[17] G. Gkioxari, R. Girshick, P. Doll´ar, and K. He, “Detecting and recognizing human-object interactions,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8359–8367.

[18] T. Golda, T. Kalb, A. Schumann, and J. Beyerer, “Human pose estimation for real-world crowded scenarios,” in 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2019, pp. 1–8.

[19] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, “Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 932–940.

[20] S. Gupta and J. Malik, “Visual semantic role labeling,” arXiv preprint arXiv:1505.04474, 2015.

[21] T. Gupta, A. Schwing, and D. Hoiem, “No-frills human-object interaction detection: Factorization, appearance and layout encodings, and training techniques,” arXiv preprint arXiv:1811.05967, 2018.

[22] Z. Hou, X. Peng, Y. Qiao, and D. Tao, “Visual compositional learning for human-object interaction detection,” arXiv preprint arXiv:2007.12407, 2020.

[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” in Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 675–678.

[24] J. Johnson, A. Gupta, and L. Fei-Fei, “Image generation from scene graphs,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1219–1228.

[25] K. Kato, Y. Li, and A. Gupta, “Compositional learning for human object interaction,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 234–251.

[26] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, “A survey of the recent architectures of deep convolutional neural networks,” arXiv preprint arXiv:1901.06032, 2019.

[27] M. Kilickaya and A. Smeulders, “Diagnosing rarity in human-object interaction detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 904–905.

[28] D. Kim, G. Lee, J. Jeong, and N. Kwak, “Tell me what they’re holding: Weakly supervised object detection with transferable knowledge from human-object interaction,” arXiv preprint arXiv:1911.08141, 2019.

[29] E. Kodirov, T. Xiang, and S. Gong, “Semantic autoencoder for zeroshot learning,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3174–3183.

[30] A. Kolesnikov, A. Kuznetsova, C. Lampert, and V. Ferrari, “Detecting visual relationships using box attention,” in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 0–0.

[31] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al., “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” International Journal of Computer Vision, vol. 123, no. 1, pp. 32–73, 2017.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.

[33] C. H. Lampert, H. Nickisch, and S. Harmeling, “Attribute-based classification for zero-shot visual object categorization,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 3, pp. 453–465, 2013.

[34] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y.-F. Wang, and C. Lu, "Transferable interactiveness prior for human-object interaction detection," arXiv preprint arXiv:1811.08264, 2018.

[35] Z. Liang, Y. Guan, and J. Rojas, "Visual-semantic graph attention network for human-object interaction detection," arXiv preprint arXiv:2001.02302, 2020.

[36] Z. Liang, J. Liu, Y. Guan, and J. Rojas, "Pose-based modular network for human-object interaction detection," arXiv preprint arXiv:2008.02042, 2020.

[37] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng, "Ppdm: Parallel point detection and matching for real-time human-object interaction detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 482–490.

[38] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick, "Microsoft coco: Common objects in context," in European conference on computer vision. Springer, 2014, pp. 740–755.

[39] T.-Y. Lin, A. Roy Chowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1449–1457.

[40] D. Liu, M. Bober, and J. Kittler, "Visual semantic information pursuit: A survey," arXiv preprint arXiv:1903.05434, 2019.

[41] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in European conference on computer vision. Springer, 2016, pp. 21–37.

[42] D. G. Lowe, "Object recognition from local scale-invariant features," in Proceedings of the seventh IEEE international conference on computer vision, vol. 2. IEEE, 1999, pp. 1150–1157.

[43] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in European Conference on Computer Vision. Springer, 2016, pp. 852–869.

[44] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in European conference on computer vision. Springer, 2016, pp. 483–499.

[45] S. Park, X. Zhang, A. Bulling, and O. Hilliges, "Learning to find eye region landmarks for remote gaze estimation in unconstrained settings," in Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications. ACM, 2018, p. 21.

[46] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in NIPS-W, 2017.

[47] J. Peyre, I. Laptev, C. Schmid, and J. Sivic, "Detecting unseen visual relations using analogies," arXiv preprint arXiv:1812.05736, 2018.

[48] J. Peyre, J. Sivic, I. Laptev, and C. Schmid, "Weakly-supervised learning of visual relations," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5179–5188.

[49] A. Prest, C. Schmid, and V. Ferrari, "Weakly supervised learning of interactions between humans and objects," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 3, pp. 601–614, 2011.

[50] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning humanobject interactions by graph parsing neural networks," in The European Conference on Computer Vision (ECCV), September 2018.

[51] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779– 788.

[52] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards realtime object detection with region proposal networks," in Advances in neural information processing systems, 2015, pp. 91–99.

[53] Y. Song, W. Li, L. Zhang, J. Yang, E. Kiciman, H. Palangi, J. Gao, C.- C. J. Kuo, and P. Zhang, "Novel human-object interaction detection via adversarial domain generalization," arXiv preprint arXiv:2005.11406, 2020.

[54] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1653–1660.

[55] R. Valenti, N. Sebe, and T. Gevers, "Combining head pose and eye location information for gaze estimation," IEEE Transactions on Image Processing, vol. 21, no. 2, pp. 802–815, 2011.

[56] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, "Pose-aware multi-level feature network for human object interaction detection," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9469–9478.

[57] S. Wang, K.-H. Yap, J. Yuan, and Y.-P. Tan, "Discovering human interactions with novel objects via zero-shot learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11 652–11 661.

[58] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional pose machines,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4724–4732.

[59] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, “A comprehensive survey on graph neural networks,” arXiv preprint arXiv:1901.00596, 2019.

[60] B. Xu, J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, “Interact as you intend: Intention-driven human-object interaction detection,” IEEE Transactions on Multimedia, 2019.

[61] B. Xu, Y. Wong, J. Li, Q. Zhao, and M. S. Kankanhalli, “Learning to detect human-object interactions with knowledge,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[62] Y. Yu, G. Liu, and J.-M. Odobez, “Deep multitask gaze estimation with a constrained landmark-gaze model,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 0–0.

[63] H. Zhang, Z. Kyaw, J. Yu, and S.-F. Chang, “Ppr-fcn: weakly supervised visual relation detection via parallel pairwise r-fcn,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4233–4241.

[64] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, “Appearance-based gaze estimation in the wild,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4511–4520.

[65] S. Zheng, S. Chen, and Q. Jin, “Skeleton-based interactive graph network for human object interaction detection,” in 2020 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2020, pp. 1–6.

[66] P. Zhou and M. Chi, “Relation parsing neural network for human object interaction detection,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 843–851.

[67] T. Zhou, W. Wang, S. Qi, H. Ling, and J. Shen, “Cascaded human object interaction recognition,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4263– 4272.

[68] B. Zhuang, Q. Wu, C. Shen, I. Reid, and A. v. d. Hengel, “Care about you: towards large-scale human-centric visual relationship detection,” arXiv preprint arXiv:1705.09892, 2017.

[69] Z. Zou, Z. Shi, Y. Guo, and J. Ye, “Object detection in 20 years: A survey,” arXiv preprint arXiv:1905.05055, 2019.