

Human-Object Interaction Detection: A Quick Survey and Examination of Methods

Trevor Bergstrom, Humphrey Shi

{bergstr,hshi3}@cs.uoregon.edu

University of Oregon

ABSTRACT

Human-object interaction detection is a relatively new task in the world of computer vision and visual semantic information extraction. With the goal of machines identifying interactions that humans perform on objects, there are many real-world use cases for the research in this field. To our knowledge, this is the first general survey of the state-of-the-art and milestone works in this field. 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. Most commonly these are the humans and objects in question, as well as a spatial quality of the two. As far as we are aware, there have not been in-depth studies performed that look into the performance of each component individually. In order to provide insight to future researchers, we perform an individualized study that examines the performance of each component of a multi-stream convolutional neural network architectures for human-object interaction detection. Specifically we examine the HORCNN architecture as it is a foundational work in the field. In addition, we provide an in-depth look at the HICO-DET dataset, a popular benchmark in the field of human-object interaction detection. Code and papers can be found at <https://github.com/SIH-Labs/Human-Object-Interaction-Detection>.

CCS CONCEPTS

- Computing methodologies → Scene understanding.

KEYWORDS

human-object interaction detection, visual relationship detection, convolutional neural networks, visual understanding

ACM Reference Format:

Trevor Bergstrom, Humphrey Shi. 2020. Human-Object Interaction Detection: A Quick Survey and Examination of Methods. In *1st International Workshop on Human-centric Multimedia Analysis (HuMA'20), October 12, 2020, Seattle, WA, USA*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3422852.3423481>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

HuMA'20, October 12, 2020, Seattle, WA, USA

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-8151-2/20/10...\$15.00

<https://doi.org/10.1145/3422852.3423481>

1 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 perform well on 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.

We choose to divide the main track of AI computer vision research into two tracks, visual perception tasks, and visual understanding tasks. Visual perception tasks focus on identifying parts or features of an image. Object detection and image segmentation are examples of well known domains that fit in the purview of visual perception. Given a scene or image, our goal is to make some quick observations that are easy to see. For example, we should be able to determine the possible objects present in an image, or whether or not a section of the image is part of an object or the background. These properties must to be learned or hand designed, but generally are not more complex than reinforcing specific feature combinations that make up a human, while a different combination of features make up a cat. Visual understanding tasks on the other hand, require far more complex analysis of a scene. These tasks focus on the less visual and at as easily recognized features of an image. Visual understanding tasks include domains such as visual relationship detection and activity recognition, as well as our focus of human-object interaction detection. In many of these tasks, visual perceptions such as object detection, are prerequisites to obtain before moving on to identifying the finer grained features needed to complete the task. For human-object interaction detection, a model must detect the possible objects in the image, then make a decision on whether an interaction is occurring and if so what that interaction is.

Human-object interaction detection is closely related to other computer vision research areas such as visual relationship detection and activity recognition. However, the foundational works and datasets differ from those used in the aforementioned tasks. We feel that there is a need for a brief and general overview for future researchers to easily obtain the baseline knowledge to create

contributions in this field. One of our main contributions is a quick survey of deep learning based methods for solving human-object interaction detection, as well as a look at some of the datasets and metrics used for evaluating models. Commonly in this field, classifying the interactions is done through the use of a multi-stream convolutional neural network. We present a detailed examination of the performance of each individual stream component. This investigation can provide useful information on how to develop future models, using this multi-stream method. Using the components from HORCNN [4], we conduct tests on their ability to correctly classify human-object interactions.

We also provide a careful analysis of the HICO-DET dataset [4]. This dataset is commonly used as a baseline for human-object interaction detection. This dataset contains a large and complex set of interactions on various objects. We provide a meaningful analysis of the various components, and suggestions for improving this state-of-the-art dataset.

2 HUMAN-OBJECT INTERACTION DETECTION DOMAIN SURVEY

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. 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. This work will focus primarily on the task of human-object interaction detection. The goal of human-object interaction detection, is to correctly identify humans, objects, and the actions that are occurring between them, if any, in an image [5]. The first step in discovering an human-object interaction from an image is to detect objects. Object proposals recovered from the image should contain at least one human for an human-object interaction to be present. Using these human and object proposals, a model for solving this problem must then correctly identify a human-object interaction between the humans present and any of the objects in the image.

We have classified the methods of solving human-object interaction 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. Graph neural networks intuitively connect objects in the image in a graphical form of nodes and connected images, that represent the relationships between objects in the image. This section will provide further insight into how each of these approaches identifies human-object interactions, as well as their strengths and drawbacks.

2.1 Multi-Stream Approaches

Multi-stream convolutional neural networks were first proposed for the task of human-object interaction detection by Chao et al. as Human-Object Region-based Convolutional Neural Networks (HORCNN) [4]. HORCNN includes three "streams", based around CNN architectures, to extract features from different sources in the image. Using object proposals from the RCNN [9] object detector, the human and object streams extract appearance queues from the image. The human stream can interpret human pose at an

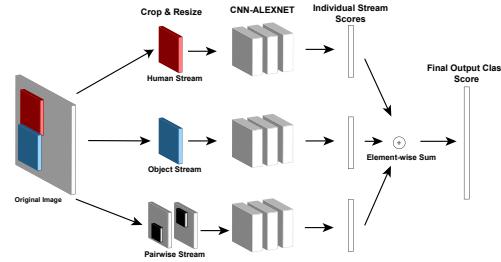


Figure 1: Diagram of the HORCNN architecture.

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 [14] implementations, pre-trained on ImageNet. At the end of each stream, classification for the possible interaction classes is performed. Finally, an element-wise sum is taken for their feature vectors for final classification scores, therefore each stream has equal an equal weight in the final classification score. 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 1.

Building on this method of multi-stream approach, Gkioxari et al. [10] use a similar architecture for detecting human-object interactions, in InteractNet. They use three branches based on Faster R-CNN feature extraction backbone: an object detection branch, a human-centric branch, and an interaction branch. The object detection branch is identical to Faster R-CNN [32], performing bounding box regression, and computing a classification score for the detected humans and objects in the image. The human centric branch performs two tasks, action classification, and target object localization. Similarly to HORCNN, human appearance features are 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, and action score to model the probability density of the target object's location. 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 performing sigmoid activation on the outputs from human action and target action classification. The final score is a product of the human and object detection scores, the target localization score, and the action classification score.

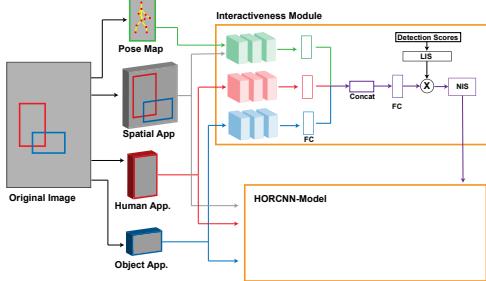


Figure 2: Diagram of the Transferable Interactiveness Network architecture.

Another implementation of the multi-stream architecture is presented by Gao et al. Instance Centric Attention Network (iCAN) [8]. They propose 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 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, 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. Scores for each action are computed similarly to InteractNet, combining the detection confidence, action, and target location probability scores.

2.1.1 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. Yao and Fei-Fei [41] very early on proposed the idea that pose information and object can provide mutual context to each other, showing that a detection of one can provide informative cues towards the detection of the other. They estimate human pose to help detect target objects for the human's interaction in the image using a random field model that learns the connectivity patterns between human body parts and objects. Fang et al. [7] propose another model that uses the individual body part attention specifically for use on human-object interaction detection datasets. The authors note that just using individual body parts 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. [12] and Li et al. [21]. Li et al. propose an add-on module to existing human-object interaction detection models, using pose information as supplemental information in their spatial information stream. The Transferable Interactiveness Network (TIN) module uses a three stream, convolutional feature extraction architecture similar to HORCNN, combining a human pose map with their spatial configuration stream. Their network works on the idea of eliminating pairs of humans and objects that are not likely to be interacting with each other. Specifically they use two functions, the low-grade suppressive (LIS) function which determines the interactiveness from the detection scores, and the non-interaction suppression function which eliminates the pairs of humans and objects that are not interacting. The final output score is incorporated into an existing model such as HORCNN as the authors used in their work. 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 Interactiveness Network module. This mechanism provides learned information from multiple human-object interaction datasets to produce a highly accurate inference on a testing image. The architecture of this add on module can be seen in Figure 3.

Another model that uses pose estimation is the Pose-aware Multi-level Feature Network or PMFNet proposed by Wan et al. [35]. This approach utilizes a slightly different architecture and score fusion 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 for feature fusion. Using human, object, and union (interaction area) proposals detected using Faster R-CNN [32] 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. [6]. 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, consisting of four streams: human, object, union, and spatial configuration. Each stream is responsible for embedding their 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, 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.

One method of note proposed by Xu et al. [39], 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 [32] is used to create human and object proposals. A pose estimation network from [6], and a gaze direction detector borrowed from [34], 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 [27], [42], [23], or [44]. 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. [47], Cascaded HOI 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 [4] and [10]. Another work called Parallel Point Detection and Matching (PPDM) [24], 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.

2.2 Graph Neural Networks

An image with human-object interactions can be interpreted similarly to a scene graph, in which the nodes represent objects and humans while the edges connecting the nodes represent relations between them. This method is very similar to the task of scene graph generation, which is followed very closely in the human-object interaction detection task by [40], breaking the task down into a graph. Qi et al. [30] propose a novel model using a graph neural network based on message passing. The goal of the model,

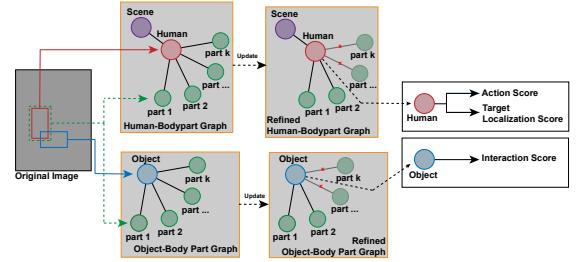


Figure 3: Diagram of the RPNN graph parsing process.

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 summarizes 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. [46] provide an improvement on the GPNN [30] 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 models body part contexts to predict actions. RPNN performs very well on HICO-DET and V-COCO. A visual illustration on how RPNN parses the two graphs can be seen in Figure 2. A more recent work into graph neural networks was conducted by Liang et. al [22], earning this paper a top mAP score for the HICO-DET dataset. Like RPNN, 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, [45] is a good example of this.

2.3 Weakly Supervised 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 have been well documented throughout the years for more classical tasks of computer vision, such as object detection [37] and segmentation [31, 38], or even without the use of deep neural networks as in [3] and [20], and using autoencoders as seen in [17].

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 [26] features or convolutional neural networks [43]. However some distribution issues can be attributed to the dataset, as seen in the study [16], 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 [28], 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 [29] 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 [2], [36], and [15]. Hou et al. [13] propose the Visual Compositional Learning (VCL) 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. In a closely related approach Bansal et al. [1] use the idea of the similarities between human actions to help guide zero-shot classes using human features from more common instances, just like the example given of human-ride-horse and human-ride-zebra. They use a word2vec model to model similarities between objects which provides a likelihood that a specific interaction can occur between the object in question and the human, based on other

Model	Default			Known Object		
	Full	Rare	Non-Rare	Full	Rare	Non-Rare
HORCNN [4]	7.81	5.37	8.54	10.41	8.94	10.85
InteractNet [10]	9.94	7.16	10.77	-	-	-
GPNN [30]	13.11	9.34	14.23	-	-	-
iCAN[8]	14.84	10.45	16.15	16.26	11.33	17.73
TIN [21]	17.03	13.42	18.11	19.17	15.51	20.26
PMFNet [35]	17.46	15.65	18.00	20.34	17.47	21.20
RPNN [46]	17.35	12.78	18.71	-	-	-
VS-GAT[22]	20.27	16.03	21.54	-	-	-
PPDM [24]	21.73	13.78	24.10	24.58	16.65	26.84
VCL [13]	23.63	17.21	25.55	25.98	19.12	28.03

Table 1: Performance of the surveyed models on HICO-DET. dashes denote un-evaluated metrics from the original work (%mAP)

similar objects. Their work shows improvement on the HICO-DET dataset and for zero-shot test cases, but they detail their approach as being limited by the fact that an interaction can look entirely different on two seemingly similar objects. Another interesting recent work on improving generalization across the lesser seen interaction examples is done by Song et al. in [33]. 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 [4] 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 [4], and using parts of the UnRel dataset [28] as a validation set. They do show great performance on zero-shot interaction categories, however we cannot rank their approach in Table 1 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 1. The scores listed were found by their authors and published in their papers. HICO-DET offers several evaluation setups and difficulties shown in this table. Table 2 contains scores for different models evaluated by average precision for a specified role in the V-COCO dataset.

2.4 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. 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

Model	<i>AP_{role}</i>
Baseline [11]	31.8
InteractNet [10]	40.0
GPNN [30]	44.0
iCAN [8]	45.3
iHOI [39]	45.9
RPNN [46]	47.5
VCL [13]	48.3
TIN [21]	48.7
Cascaded HOI [47]	48.9
VS-GAT[22]	49.8
PMFNet [35]	52.0

Table 2: Performance of the surveyed models on V-COCO. (%AP)

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 3 summarizes the datasets and their properties, as discussed in this section.

One of the first purpose-built datasets for the task of human-object interaction detection is the Humans Interacting with Common Objects (HICO) [5] dataset, created by Chao et al. This dataset was constructed from the MS-COCO [25] 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. HICO 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 an 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 with Detection (HICO-DET) [4]. 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 [11], is another commonly evaluated dataset for human-object interaction detection. Similar to HICO, the object classes are taken from the COCO [25] 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. [48]. 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 [18], 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 [5]. 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. Testing on the HICO and HICO-DET datasets is 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 [11]. 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. 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 [50]. 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 [28]. 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 [33] 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.

3 PROPOSED WORK

To understand and study the multi-stream architectures, we propose to implement the HORCNN detection model using the PyTorch

	<i>Images</i>	<i>Interaction Classes</i>	<i>Object Classes</i>
HICO [5]	47,774	600	80
HICO-DET [4]	47,776	600	80
V-COCO [11]	10,346	26	80
HCVRD [48]	52,855	927	1824

Table 3: human-object interaction Dataset Metrics.

deep learning framework. 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 individual RCNN [9] detectors for each object for their object detectors. We chose to use Faster-RCNN [32] to its immediate availability as a module included with PyTorch. In this implementation, Faster-RCNN is pre-trained on the MS-COCO dataset which includes the same object classes as HICO-DET. Another deviation is the use of AlexNet [19] rather than CaffeNet [14]. For their implementation of HORCNN, the authors use CaffeNet pre-trained on the ImageNet dataset for the human and object convolutional streams. 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. AlexNet and CaffeNet are the same architecture, with AlexNet being modified for multi GPU use. With our implemented version of the HORCNN model, we will follow the training process as detailed in the original work. With this trained model, we will perform an extensive ablation study where we break the model into its various streams and combinations. With the results found in this study we should be able to see which model components and which image features give the best performance for to HICO-DET dataset. Our prediction is that the human centric component of the network should provide the model the most detailed visual features for predicting a human-object interaction class. This is driven by the understanding that a human should exhibit the intent to interact on the object. For example, features extracted from a human kicking a sports ball should be far different than features extracted from a human throwing a sports ball. From the other end, we should see that the features extracted from a sports ball being kicked should remain similar to those of a sports ball being thrown.

3.1 Experiments and Results

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 II 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

<i>Implementation</i>	<i>Full</i>	<i>Rare</i>	<i>Non-Rare</i>
Ours	5.87	3.06	7.08
Chao et al. [4]	7.81	5.37	8.54

Table 4: Performance (%mAP) of re-implemented model vs. published results

<i>Implementation</i>	<i>Full</i>	<i>Rare</i>	<i>Non-Rare</i>
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
H, O	5.41	3.51	6.54
H, P	1.41	0.15	1.65

Table 5: Performance of individual model streams (%mAP).
H, O,P denote Human, Object, and Pairwise streams respectively

iterations at a learning rate of 0.0001. Results and comparisons can be seen in Table 4 and Table 5. The model was trained for 20 hours on a single Nvidia TitanXp GPU.

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. Results of these tests can be seen in Table 5. 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. 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,

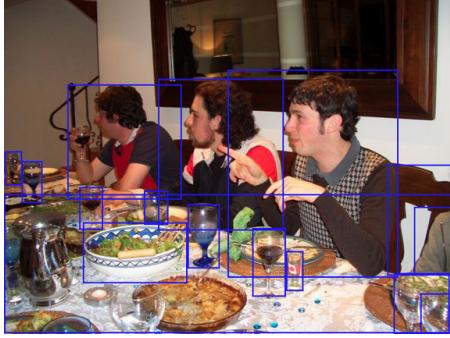


Figure 4: Example of non-exhaustively labeled image from HICO-DET

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.

3.2 Dataset Discussion

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 instances in the images where there should be a no interaction category, but they are not labeled 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 is taken from the HICO-DET training set with bounding boxes representing detections from FastRCNN. 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”.

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



Figure 5: Examples of the interaction class ‘human repair mouse’ from the HICO-DET dataset.

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.

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 5 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, these are the only 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.

3.3 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.

REFERENCES

- [1] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. 2020. Detecting Human-Object Interactions via Functional Generalization. In *AAAI*. 10460–10469.
- [2] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. 2018. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 384–400.
- [3] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. 2015. Weakly supervised object detection with convex clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1081–1089.
- [4] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. 2018. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*. IEEE, 381–389.
- [5] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. 2015. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*. 1017–1025.
- [6] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. 2018. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7103–7112.
- [7] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. 2018. Pairwise body-part attention for recognizing human-object interactions. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 51–67.
- [8] Chen Gao, Yuliang Zou, and Jia-Bin Huang. 2018. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437* (2018).
- [9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.
- [10] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. 2018. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8359–8367.
- [11] Saurabh Gupta and Jitendra Malik. 2015. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474* (2015).
- [12] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. 2018. No-Frills Human-Object Interaction Detection: Factorization, Appearance and Layout Encodings, and Training Techniques. *arXiv preprint arXiv:1811.05967* (2018).
- [13] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. 2020. Visual Compositional Learning for Human-Object Interaction Detection. *arXiv preprint arXiv:2007.12407* (2020).
- [14] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*. 675–678.
- [15] Keizo Kato, Yin Li, and Abhinav Gupta. 2018. Compositional learning for human object interaction. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 234–251.
- [16] Mert Kilicaya and Arnold Smeulders. 2020. Diagnosing Rarity in Human-Object Interaction Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 904–905.
- [17] Elyor Kodirov, Tao Xiang, and Shaogang Gong. 2017. Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3174–3183.
- [18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.
- [19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.
- [20] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2013. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 3 (2013), 453–465.
- [21] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yan-Feng Wang, and Cewu Lu. 2018. Transferable interactiveness prior for human-object interaction detection. *arXiv preprint arXiv:1811.08264* (2018).
- [22] Zhijun Liang, Yisheng Guan, and Juan Rojas. 2020. Visual-Semantic Graph Attention Network for Human-Object Interaction Detection. *arXiv preprint arXiv:2001.02302* (2020).
- [23] Zhijun Liang, Junfa Liu, Yisheng Guan, and Juan Rojas. 2020. Pose-based Modular Network for Human-Object Interaction Detection. *arXiv preprint arXiv:2008.02042* (2020).
- [24] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. 2020. 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*. 482–490.
- [25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.
- [26] David G Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, Vol. 2. Ieee, 1150–1157.
- [27] Seonwook Park, Xucong Zhang, Andreas Bulling, and Otmar Hilliges. 2018. 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, 21.
- [28] Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2017. Weakly-supervised learning of visual relations. In *Proceedings of the IEEE International Conference on Computer Vision*. 5179–5188.
- [29] Alessandro Prest, Cordelia Schmid, and Vittorio Ferrari. 2011. Weakly supervised learning of interactions between humans and objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 3 (2011), 601–614.
- [30] Siyuan Qi, Wenguan Wang, Baoxiang Jia, Jianbing Shen, and Song-Chun Zhu. 2018. Learning Human-Object Interactions by Graph Parsing Neural Networks. In *The European Conference on Computer Vision (ECCV)*.
- [31] Rui Qian, Yunchao Wei, Honghui Shi, Jiachen Li, Jiaying Liu, and Thomas Huang. 2019. Weakly supervised scene parsing with point-based distance metric learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8843–8850.
- [32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.
- [33] Yuhang Song, Wenbo Li, Lei Zhang, Jianwei Yang, Emre Kiciman, Hamid Palangi, Jianfeng Gao, C-C Jay Kuo, and Pengchuan Zhang. 2020. Novel Human-Object Interaction Detection via Adversarial Domain Generalization. *arXiv preprint arXiv:2005.11406* (2020).
- [34] Roberto Valenti, Nicu Sebe, and Theo Gevers. 2011. Combining head pose and eye location information for gaze estimation. *IEEE Transactions on Image Processing* 21, 2 (2011), 802–815.
- [35] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. 2019. Pose-aware Multi-level Feature Network for Human Object Interaction Detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 9469–9478.
- [36] Suchen Wang, Kim-Hui Yap, Junsong Yuan, and Yap-Peng Tan. 2020. Discovering Human Interactions with Novel Objects via Zero-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11652–11661.
- [37] Yunchao Wei, Zhiqiang Shen, Bowen Cheng, Honghui Shi, Jinjun Xiong, Jiashi Feng, and Thomas Huang. 2018. Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 434–450.
- [38] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. 2018. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7268–7277.
- [39] Bingjie Xu, Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. 2019. Interact as you intend: Intention-driven human-object interaction detection. *IEEE Transactions on Multimedia* (2019).
- [40] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. 2019. Learning to detect human-object interactions with knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- [41] Bangpeng Yao and Li Fei-Fei. 2010. Modeling mutual context of object and human pose in human-object interaction activities. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 17–24.
- [42] Yu Yu, Gang Liu, and Jean-Marc Odobez. 2018. Deep multitask gaze estimation with a constrained landmark-gaze model. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 0–0.
- [43] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. 2017. PPR-FCN: weakly supervised visual relation detection via parallel pairwise R-FCN. In *Proceedings of the IEEE International Conference on Computer Vision*. 4233–4241.
- [44] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4511–4520.
- [45] Sipeng Zheng, Shizhe Chen, and Qin Jin. 2020. Skeleton-Based Interactive Graph Network For Human Object Interaction Detection. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.
- [46] Penghao Zhou and Mingmin Chi. 2019. Relation parsing neural network for human-object interaction detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 843–851.
- [47] Tianfei Zhou, Wenguan Wang, Siyuan Qi, Haibin Ling, and Jianbing Shen. 2020. Cascaded human-object interaction recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4263–4272.
- [48] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton van den Hengel. 2018. Hcvrd: a benchmark for large-scale human-centered visual relationship detection. In *Thirty-Second AAAI Conference on Artificial Intelligence*.