Scaling Human-Object Interaction Recognition through Zero-Shot Learning

Liyue Shen
Stanford University
liyues@stanford.edu

Serena Yeung
Stanford University
serena@cs.stanford.edu

Judy Hoffman
UC Berkeley*
jhoffman@eecs.berkeley.edu

Greg Mori
Simon Fraser University
mori@cs.sfu.ca

Li Fei-Fei
Stanford University
feifeili@cs.stanford.edu

Abstract

Recognizing human object interactions (HOI) is an important part of distinguishing the rich variety of human action in the visual world. While recent progress has been made in improving HOI recognition in the fully supervised setting, the space of possible human-object interactions is large and it is impractical to obtain labeled training data for all interactions of interest. In this work, we tackle the challenge of scaling HOI recognition to the long tail of categories through a zero-shot learning approach. We introduce a factorized model for HOI detection that disentangles reasoning on verbs and objects, and at test-time can therefore produce detections for novel verb-object pairs. We present experiments on the recently introduced large-scale HICO-DET dataset, and show that our model is able to both perform comparably to state-of-the-art in fully-supervised HOI detection, while simultaneously achieving effective zero-shot detection of new HOI categories.

1. Introduction

Humans are a main focus in the visual world, and as such a core challenge in computer vision is distinguishing the rich variety of human actions. Key to this problem is recognizing human object interactions, since many action categories are defined by subtle differences between these interactions. For example, "riding a bike" and "walking a bike" are often distinct categories of interest despite their similar visual appearance. This has led to recent interest and work addressing human-object interaction (HOI) recognition [26, 6, 19, 22, 7, 15, 5, 20, 27, 12], and datasets to support the task [17, 5, 27, 14]. In particular, the recent "Humans Interacting with Common Objects" dataset for classification (HICO) [5] and detection (HICO-DET) [27] represents a challenging new HOI benchmark at substantially larger scale.

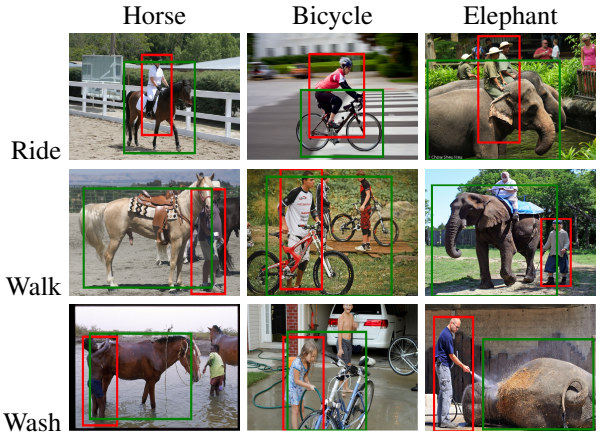

Figure 1: To scale to all combinations of human-object interactions we use a multi-task approach. Each row shows example detections for the same verb. Each column shows example detections for the same object.

When considering the problem of HOI understanding, an important desideratum is the ability to scale recognition to the long tail of HOI categories. Given the huge space of possible human-object interactions, it is impractical to obtain labeled training data for all interactions of interest. Recent work in HOI understanding has focused primarily on improving recognition performance through approaches such as improved feature representations [6, 19, 5, 20, 27] or modeling of spatial interaction and context [26, 7, 27]. However, these methods are all fully-supervised and limited to recognition of classes for which labeled training data is available.

A promising approach to address the problem of scaling HOI recognition to the long tail of classes is zero-shot learning. In zero-shot learning, recognition of previously unseen classes is accomplished through knowledge learned from

*Work done while a Stanford Postdoc.

training data of other classes. Approaches for this include metric learning, attribute recognition, and domain transfer-based methods [3, 28, 8, 16, 1, 2, 25, 21]. These have been successfully applied for tasks such as image and scene classification, and fine-grained bird classification. However, zero-shot learning has not yet been investigated in the context of human-object interaction recognition, where a structured relationship exists between humans and objects.

In this work, we address the challenge of scaling human-object interaction recognition by introducing an approach for zero-shot learning that reasons on the decomposition of HOIs as verbs and objects. Specifically, we tackle the problem of HOI detection, and introduce a factorized model consisting of both shared neural network layers as well as independent verb and object networks. The entire model is trained jointly in a multi-task fashion, but produces disentangled verb and object networks that can be used at test time to recognize novel verb-object pairs based on previously seen instances of the verb or object. We perform experiments on the HICO-DET dataset, and demonstrate that our factorized model is able to perform both comparably to state-of-the-art in fully-supervised HOI detection, as well as effectively detect novel HOI categories.

2. Related work

Human-Object Interaction. There has been a recent body of work on modeling human-object interactions (HOI) in images [26, 6, 19, 22, 7, 15, 5, 20, 27]. Yao et al. [26] uses a random field model to encode mutual context of human pose and objects. Delaitre et al. [6] introduces interaction features to model spatial relationships of humans and objects, while Hu et al. [15] takes an exemplar-based approach. Maji et al. [19] and Desai et al. [7] learn distributed representations of human and object in the form of poselets [19] and relational phraselets [7]. More recently, Chao et al. [5] introduced a new benchmark, "Humans Interacting with Common Objects" (HICO), for HOI recognition, which was expanded for detection in HICO-DET [27]. This is the first large scale dataset for HOI recognition, with 150K instances of 600 HOI categories. [20] extracts CNN-based appearance features from human and object detections to obtain state-of-the-art results on recognition, while [27] also uses a human and object detector-based approach, combined with spatial relationship features, for detection.

In contrast to these works, which focus on building stronger recognition models for fully-supervised HOI prediction, our work addresses the task of zero-shot learning for HOI. To the best of our knowledge, we are the first to introduce an approach for this problem, which enables scaling recognition to the long-tail of HOI categories.

Object Detection. Our work focuses on HOI recognition in the detection setting, which allows spatial localization of multiple HOIs per image. While this has only recently been explored for HOI [27], a large body of work has studied detection for objects in images. Recently, [11, 9, 23] use region-based convolutional neural network approaches to achieve state-of-the-art results in object detection. Our approach for HOI detection leverages the Faster R-CNN network [23] but incorporates it into a factorized model for joint verb and object detection in images.

Zero-Shot Learning. A variety of approaches have been introduced for zero-shot learning [3, 28, 8, 16, 13, 1, 2, 25, 21, 18], including metric learning, attribute recognition, and domain transfer-based methods. However, none of these methods address the problem of human-object interaction recognition, which involves multiple interacting components. Our approach of learning a factorized model of verbs and objects is most related to attribute-based methods [1, 16, 2]. Similar to these, we reason on new, unseen classes based on semantic subcomponents; in contrast, we model structured relationships between humans and interacting objects.

3. Approach

We study the problem of human-object interaction (HOI) detection in the zero-shot setting. In HOI detection, the input is an image I, and the output is a set of possibly multiple detected HOI categories and their spatial regions R_i.

An HOI category is defined as an action (verb) and object pair, $\{v_i, o_i\}$. Let us denote the set of possible verbs as $\mathcal{V}$ and the set of possible objects of interaction as $\mathcal{O}$. Direct supervised learning to produce verb-object pairs would require annotations for $|\mathcal{V}| \cdot |\mathcal{O}|$ categories. However, such an approach is actually unnecessarily redundant, since many actions involve the same object of interaction, e.g. washing or riding a bike, or the same human action, e.g. feeding a horse or a dog.

Inspired by this observation, we introduce a factorized model that is parametrized by human action (verb) prediction and object prediction. By disentangling the two components of reasoning, we remove the necessity of annotations for all $|\mathcal{V}| \cdot |\mathcal{O}|$ categories, and instead enable full pairwise prediction with annotations for only $|\mathcal{V}| + |\mathcal{O}|$ categories. Following, we describe our model architecture and training in greater detail.

3.1. Model Architecture

Our model architecture (Fig. 2) consists of a common trunk of visual feature extraction layers, followed by disentangled verb detection and object detection networks. In this way, we explicitly model and learn representations for both verbs and objects, which can later be combined in different pairings for zero-shot learning. The full model is trained end-to-end with a multi-task objective for the verb

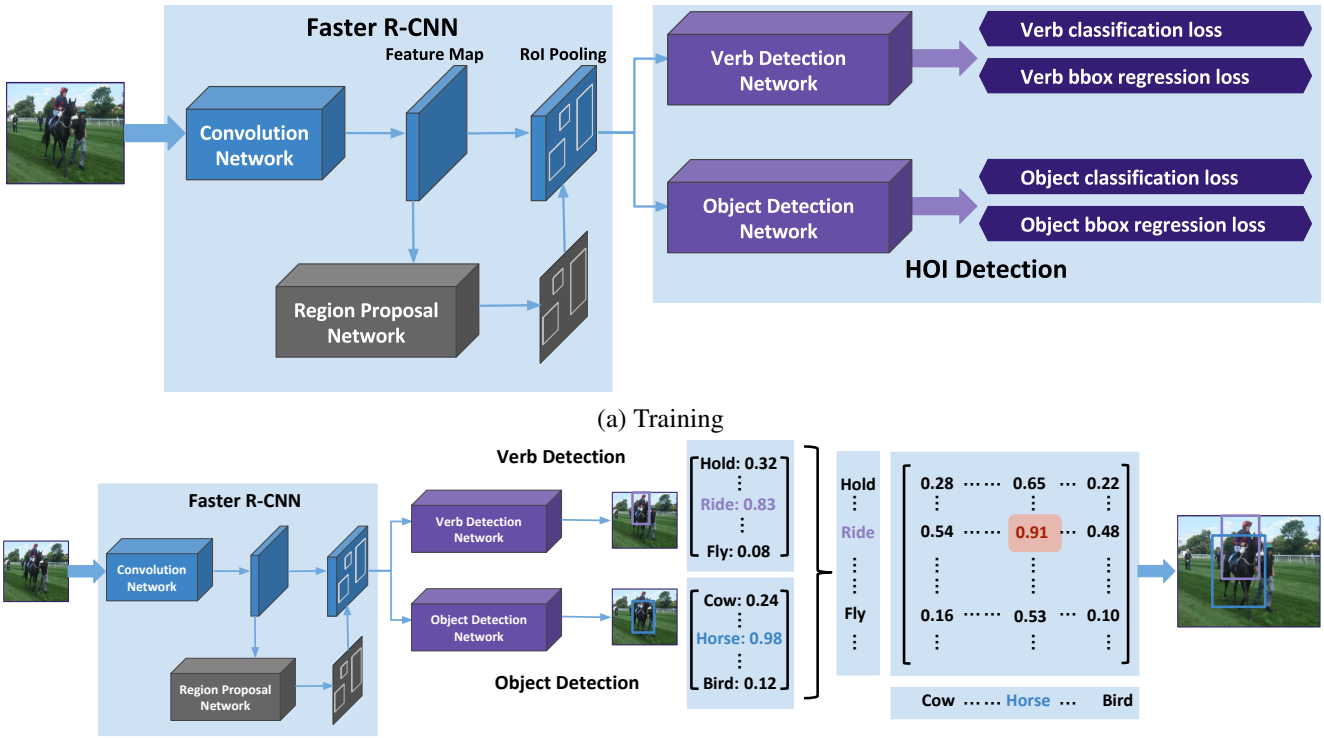

Figure 2: Overview of our model. We propose a multi-task learning framework for detection of human-object interactions (HOIs). By independently expressing the verb and object learning objectives our model is able to produce novel human-object interaction predictions at test time. The full joint model is trained end-to-end to produce predicted bounding boxes and scores for all verb and all object classes. At test time we compute scores for all combinations of verb-object prediction pairs to produce the final HOI prediction where the verb and object are tightly localized.

and object pathways. Our approach contrasts with previous deep learning-based approaches for reasoning independently on humans and objects [5, 20, 27], where appearance and pose-based features are simply aggregated over pre-detected human and object regions and input into further processing for direct HOI prediction.

The input to our model is an image containing potentially multiple human-object interactions. The output is spatial detections of human-object interactions consisting of verb-object pairs and the respective bounding boxes of both verb and object. Each input image is first passed through a common CNN feature extraction trunk consisting of 5 convolution layers and ReLU layers following the convolutional structure of VGG-19 architecture [24]. This produces a set of feature maps $\{F_i\}$, which are then passed to the disentangled verb prediction network (top stream) and object prediction network (bottom stream). At training time (Fig. 2 (a)), a multi-task objective for verb prediction and object prediction is used to jointly train the two streams of the model. At test time (Fig. 2 (b)), the verb and object prediction networks can be used independently for prediction of both previously seen verb-object pairs (standard supervised recognition) as well as new verb-object combination pairs (zero-shot recognition). Following, we describe the verb and object prediction networks in more detail.

3.1.1 Verb Network

While region-based appearance features are effective for many recognition tasks, human actions (verbs) are often expressed in part through subtle differences in body position. This subtlety is challenging, but also enables leveraging structure through reasoning on relationships between body joints (pose). We therefore base our verb network (top branch of Fig. 2) on both appearance features as well as human pose features.

Appearance Feature. We use the Region Proposal Network (RPN) from [23] to determine human locations of candidate verb region over which to pool appearance features. The RPN proposes a collection of possible bounding boxes $\{B_j\}$ from the full image, and then projects $\{B_j\}$ onto appearance feature maps $\{F_i\}$ from the common trunk to extract corresponding Regions of Interests (RoI) $\{R(F_i, B_j)\}$. A RoI pooling layer pools over the RoIs to produce a feature representation $\{\widehat{R}(F_i, B_j)\}$ per RoI.

These are passed through two fully-connected layers to produce the verb-appearance feature.

Pose Feature. The pose feature is extracted from the part affinity fields-based pose estimation network of [4], pooled over the same RoIs described above. The network consists of 6 stages, where each stage consists of 7 convolutional layers and 6 ReLU layers. The output of this network is a set of pose heatmaps $\{H_i\}$, where each heatmap is the probability distribution of a human joint's location in the image. In total, there are 18 heatmaps for the set of joints $\mathcal{J}$ = {"nose", "right shoulder", "right elbow", "right wrist", "left shoulder", "left elbow", "left wrist", "right hip", "right knee", "right ankle", "left hip", "left knee", "left ankle", "left eye", "right eye", "left ear", "right ear"}. These heatmaps are then passed through two fully-connected layers to produce the verb-pose feature.

Finally, the output of the verb network is the joint feature $\widehat{R}_{joint}(i,j) = [\widehat{R}(F_i, B_j), \widehat{R}(H_i, B_j)]$, which is a concatenation of the verb-appearance and verb-pose features. This provides a rich description of body movement.

3.1.2 Object Network

For the object network (bottom branch of Fig. 2), we leverage the framework of Faster R-CNN [23], which has shown state-of-the-art performance on object detection tasks. As in the verb network, we first use a Region Proposal Network (RPN) to determine candidate object locations, and then use a RoI layer to pool appearance features over these locations. In this network, we use only appearance features, and the object feature is produced through two fully-connected layers after the RoI layer. This feature is then used for the object classifier.

3.2. Training

Our full model is trained jointly in an end-to-end fashion using equally weighted multi-task learning objectives. Our model optimizes six different objectives. Following the Faster R-CNN model [23], we use a region proposal network (RPN) which is trained to generically produce regions likely to contain objects. The RPN has two direct loss objectives during training. First, the objectness loss is a standard softmax loss over two classes of whether the box contains an object or not. We assign positive labels for all boxes which overlap with either the ground truth object or verb boxes by more than 0.7. Second, the Smooth L1 regression loss for bounding box regression between the anchor box and the ground truth box.

The verb and object branches of our model are each trained with two independent objectives, shown in Fig. 2 (a). Independent sigmoid cross-entropy losses for both verb or object categories, respectively, and a bounding box regression loss. Only bounding boxes positive for the respective branch (verb or object) are considered for regression training using the respective verb or object Smooth L1 loss.

We initialize our model using the released COCO-VGG19 Faster R-CNN weights for the object and verb appearance networks. We initialize the pose feature network with the released weights of [4] through to pose heatmaps and randomly initialize the two additional fully connected layers. All output layers are randomly initialized.

Our model is trained using stochastic gradient descent with a fixed learning rate of 0.001 and momentum of 0.9. Following the protocol introduced in Faster R-CNN [23] we fix the first two convolutional layers of the trunk network. Additionally, in our experiments we fix the parameters corresponding to the pose heatmap predictions as the HICO-DET dataset lacks pose annotations. All other parameters are learned simultaneously through joint fine-tuning.

3.3. Testing

At test time, our model produces zero-shot HOI detections in the following manner, illustrated in Fig. 2 (b). An image is passed through the shared network layers, followed by the disentangled verb and object networks. Due to the training procedure, the outputs of these networks now produce direct prediction probabilities of each of the verb classes and object classes seen in at least one HOI verb-object pair during training. These outputs are then linked to produce a $|\mathcal{V}| \times |\mathcal{O}|$ matrix P of HOI predictions, where $|\mathcal{V}|$ is the number of verb classes seen during training and $|\mathcal{O}|$ is the number of object classes seen during training. Each element $P_{i,j}$ of P represents the linking of verb detection i with object detection j, and HOI score $P_{i,j}$ is the average of the verb detection score and object detection score.

To obtain the final HOI predictions for an image, we keep only verb-object pairs that have spatially close relation. Specifically, we keep pairs where the verb and object detections overlap on at least one axis. In this way, we are able to learn and then re-purpose our verb and object detection networks to detect up to $|\mathcal{V}| \times |\mathcal{O}|$ HOI classes, despite requiring training data for only $|\mathcal{V}| + |\mathcal{O}|$ classes at minimum.

4. Experiments

This section presents experiments on using our model for zero-shot detection of human-object interactions, based on the "Humans Interacting with Common Objects Detection" (HICO-DET) dataset [27]. We train our model using a subset of HOI classes, and evaluate on the remaining subset. HICO-DET is the first large-scale HOI dataset that enables the study of zero-shot HOI detection.

Following, we first describe the HICO-DET dataset. We then present experiments for our model on fully supervised

HOI detection, and show comparable performance on this task with state-of-the-art approaches designed for the setting. Finally, we present quantitative and qualitative results demonstrating the effectiveness of our method for zero-shot detection.

4.1. HICO-DET Dataset

The "Humans Interacting with Common Objects" (HICO) dataset [5] is a recently introduced dataset for human-object interaction (HOI) that is the first large-scale benchmark for this task. The HICO-DET dataset [27] is an extension to HICO containing spatial annotations of the humans and objects that comprise each HOI, enabling work on detection of possibly multiple HOIs in an image.

HICO-DET contains a total of 47,774 images with annotations of 600 HOI categories. HOI categories (e.g. "ride-bike", "walk-bike", "ride-horse") span 117 action categories (e.g. "ride", "walk") and 80 objects (e.g. "bike", "horse"). These describe diverse interactions per object category, with an average of 6.5 action categories associated with each object. Each image can contain possibly multiple HOI annotations, with 150K annotations total. 38116 images (80%) are used for training, and 9958 images (20%) are used for testing.

We note that HICO-DET detections are incomplete; since it is extremely label intensive to label instances of 600 HOI categories, annotations are grouped into four types: "verified positives", "verified negatives", "ambiguous/uncertain", and "unknown", depending on whether an HOI class is verified to be in the image, not in the image, annotated with disagreement, or unknown. This highlights the challenge of scaling HOI recognition and the benefits of zero-shot learning.

4.2. Analysis under Fully-Supervised Setting

All prior work on HOI detection has targeted the fully-supervised setting. While our model is designed for the zero-shot setting, it can also be evaluated on fully supervised HOI detection. To do this, we use the same verb-object linking procedure as in our zero-shot process to produce detections of training classes. We therefore first analyze results of our model under this setting to compare the representation ability of our model and its components.

Overall Detection AP. Table 1 shows fully-supervised detection mAP of our model compared to previous work. Results are separated into mAP computed using all categories ("full"), rare categories with less than 10 training images ("rare"), and non-rare categories with more than 10 training images ("non-rare"). We observe that overall, and across all types of categories, our approach outperforms or performs comparably with methods trained specifically for fully-supervised detection. In particular, we outperform the HO method [27], which detects humans and objects using a pre-trained R-CNN detector, pools local features over these regions, and trains a binary classifier for each HOI class.

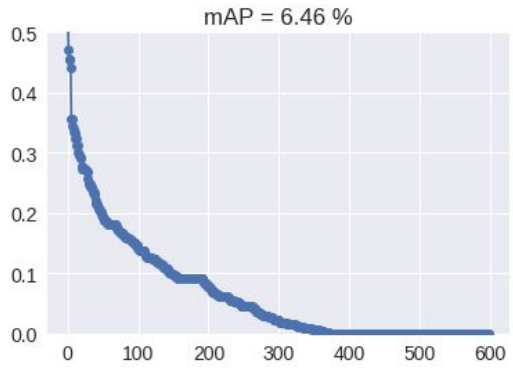

Figure 3: Sorted per-class detection AP in the fully-supervised setting. Since certain objects or verbs are more prevalent even across distinct HOI classes, we observe that per-class performance also follows a long-tail distribution.

The only models ours does not outperform are the HO+IPI - based models. These models augment the standard R-CNN architecture by inputting maps of computed spatial interaction into the 600-way classification network. This design of spatial modeling is orthogonal but complementary to work on zero-shot recognition and could be combined in future work; however, we note that even without this our model is able to perform comparably.

Ablation Studies. We analyze the effect of various components of our model on detection performance. In particular, Table 1 shows detection results of our model without pose component of the verb network, or without multi-task training of the whole network. Our full model outperforms the model without pose features, highlighting the value of reasoning on structured pose information for distinguishing subtle differences in how humans interact with objects. Additionally, based on the model without multi-task performance, multi-task training turns out to be critically important to the overall performance of our method. This indicates that information shared between the verb and object models improves individual prediction performance.

Per-Class AP. We visualize the per HOI class average precision performance in Fig. 3. Instance annotations are available in a long tailed distribution. Since certain objects or verbs are more prevalent even across distinct HOI classes, we observe that our per-class performance also follows a long-tail distribution. The 5 top performing classes are "hose elephant", "jump horse", "inspect fire hydrant", "stand under stop sign", and "milk cow". These follow intuition that classes with distinctive objects and human poses are easiest to recognize. On the other hand, some of the lowest-performing classes are "cook carrot", "control tv", "wash sheep", "tag person", and "wash fork", which again make sense given the subtle interactions and small objects

Method	Full	Rare	Non-Rare
Random	1.35×10^{-3}	5.72×10^{-4}	1.62×10^{-3}
Fast-RCNN [10] (union)	1.75	0.58	2.10
Fast-RCNN [10] (score)	2.85	1.55	3.23
HO [27]	5.73	3.21	6.48
Ours - w/o multi-task training	3.38	3.19	3.44
Ours - w/o pose	5.62	**4.37**	6.00
Ours	**6.46**	4.24	**7.12**
HO+IPI (conv) [27]	7.30	4.68	8.08
HO+IPI (conv)+S [27]	7.81	5.37	8.54

Table 1: Comparison of HOI prediction mAP(%) across 600 HOI classes in HICO-DET dataset. Our method outperforms prior models when comparing base architecture performance. The full HO+IPI [27] model incorporating additional logic to combine objects and verbs into HOI predictions is orthogonal to the base model.

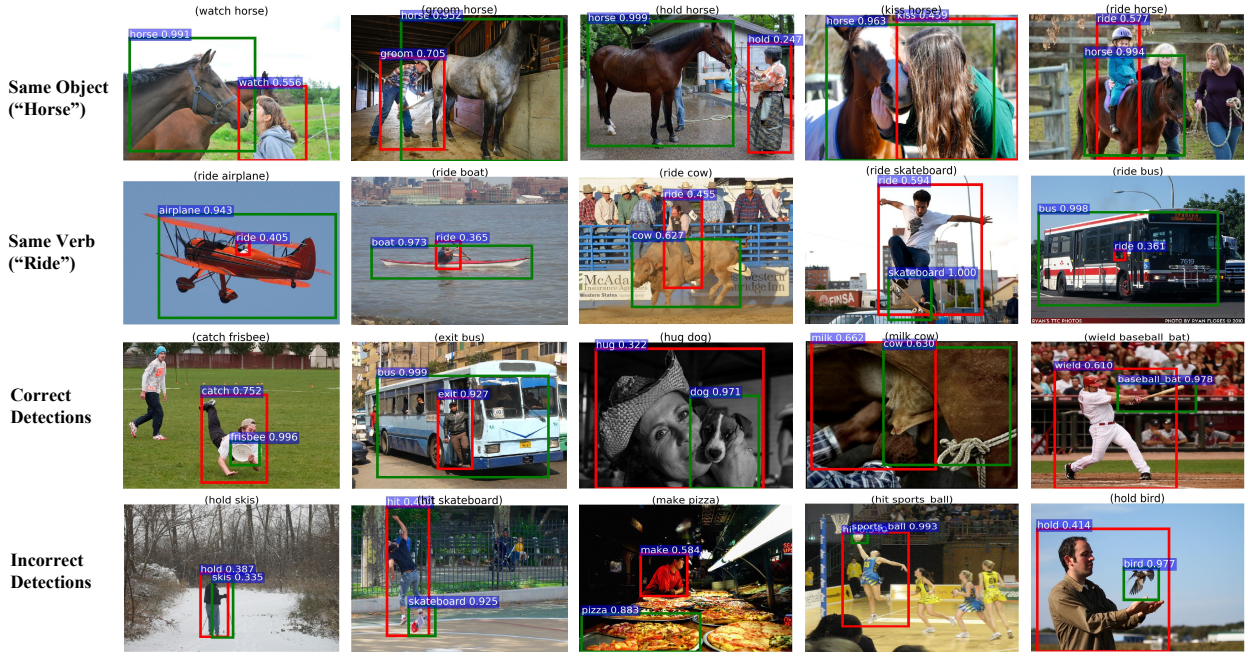

Figure 4: Example detections from our full jointly trained model in the fully-supervised setting. In the first two rows we show diverse examples for a shared object class (*top row*) or a shared verb class (*second row*). In the bottom two rows we present randomly sampled examples both where our model predicts the correct verb-object pair (*third row*) and incorrect verb-object pair (*bottom row*).

involved.

Qualitative Results. In Fig. 4 we present qualitative results of our model for the fully-supervised training setting. In the first two rows we show diverse examples for a shared object class (*top row*) or a shared verb class (*second row*). In the bottom two rows we present randomly sampled examples both where our model predicts the correct verb-object pair (*third row*) and incorrectly predicts the verb-object pair (*bottom row*). We find that our false positive detections tend to be fairly reasonable. For example the detection in the bottom right corner reports the verb "hold" which is incorrect since the bird has technically taken flight. However, we can understand why the model would predict this verb since the man was likely just previously holding the bird and is still holding his arms in a pose indicative of the "hold" verb. Similarly the person in the bottom second from the left is jumping with an outstretched arm so a prediction of "hit" is a reasonable mistake.

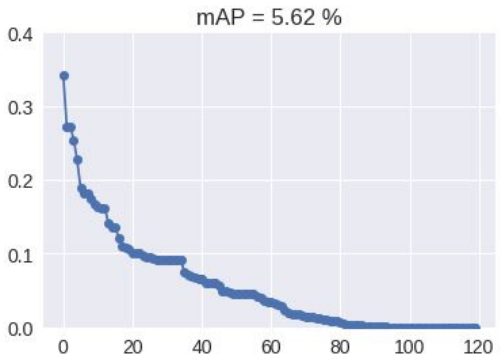

Figure 5: Sorted per-class detection AP in the zero-shot setting. The distribution is similar to the fully-supervised setting and performance is close overall.

4.3. Zero-Shot Recognition

In this section, we present results of our model for the zero-shot recognition task. For this task, we split the 600 HOI categories in HICO-DET into 480 that we use for training, and 120 for testing. We randomly choose 120 categories, ensuring only that every verb or object within the 120 categories shows up at least once in the training categories. We train on the subset of HICO-DET training images which contains the 480 training classes in our experiment. We evaluate on the HICO-DET test images, computing AP and mAP for the subset of 120 zero-shot classes not seen during training.

Quantitative Analysis. Fig. 5 shows per-class AP of our model on the 120 zero-shot classes. The overall mAP is 5.62. This is only slightly lower than the mAP for the fully-supervised setting, indicating that our model has learned robust representations of decomposed verb and object that it can effectively re-purpose to detect novel verb-object pairs. From Fig. 5, we also observe that the highest-performing classes achieve around 0.35 AP, and the distribution has a similar shape to the fully-supervised setting, with just a slight decrease in performance overall. If we evaluate our model on the full set of test classes (including the 480 training classes as well as the 120 zero-shot classes), we achieve a mAP of 6.26. This is only slightly lower than the 6.46 mAP of our fully-supervised model, showing that zero-shot learning only slightly reduced performance and the prediction ability is similar to if we had full annotations for all classes.

Looking at individual class performance of the zero-shot model, the top performing classes are "wash train", "eat orange", "wash knife", "ride horse", and "kiss elephant", while the worst-performing classes are "hug suitcase", "inspect sports ball", "wash cat", "inspect dog", and "lick person". Many of these again follow intuition: obvious objects and verbs such as train and horse, and riding and washing,

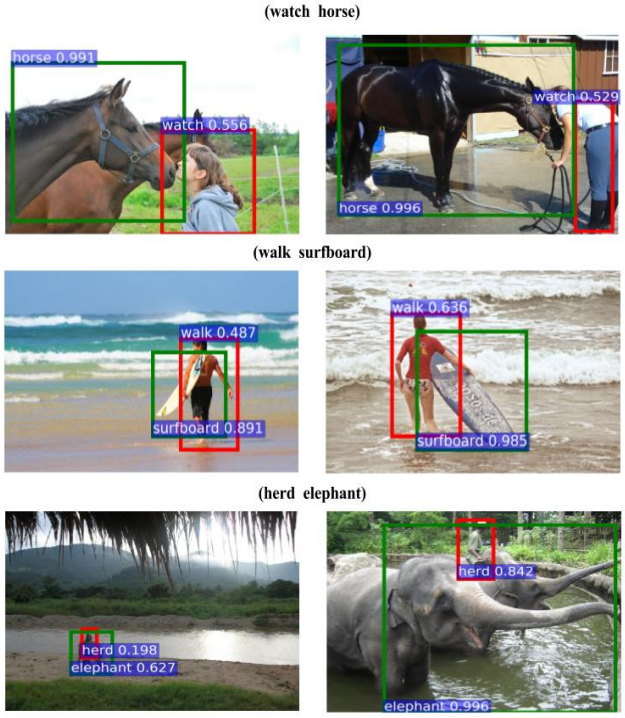

Figure 6: Our zero-shot learning algorithm discovers novel verb-object pairs at test time. These pairs are absent from the annotated set available within the HICO-DET dataset. We present here a handful of manually verified examples. For example, we previously learned the verb "herd" and to recognize elephants. Together we are able to recognize the novel pairing of herding elephants.

are easier than subtle or rare classes such as inspecting a dog or licking a person. However it is also interesting to note that the zero-shot model actually performs better on some rarer classes such as eating an orange or washing a knife, since it does not overfit to the few labeled examples present at training-time and instead uses only the more robust learned representations of the component verb and object.

Qualitative Analysis. Fig. 7 shows qualitative examples of zero-shot detections from our model. We can see that the model has learned to compose novel combinations of objects and verbs as a result of our multi-task learning architecture. For example, the model has learned the concept of a hug from training examples and can apply this to detect hugs of other known entities such as sheep and dog ("hug sheep" and "hug dog" in the right-most images of the first and second rows of Fig. 7. Other examples show that the model has also learned the concepts of holding something, washing something and watching something, and can transfer this to known objects such as skateboards, surfboards, airplanes, and cows.

Most impressively, our model even discovers new verb-

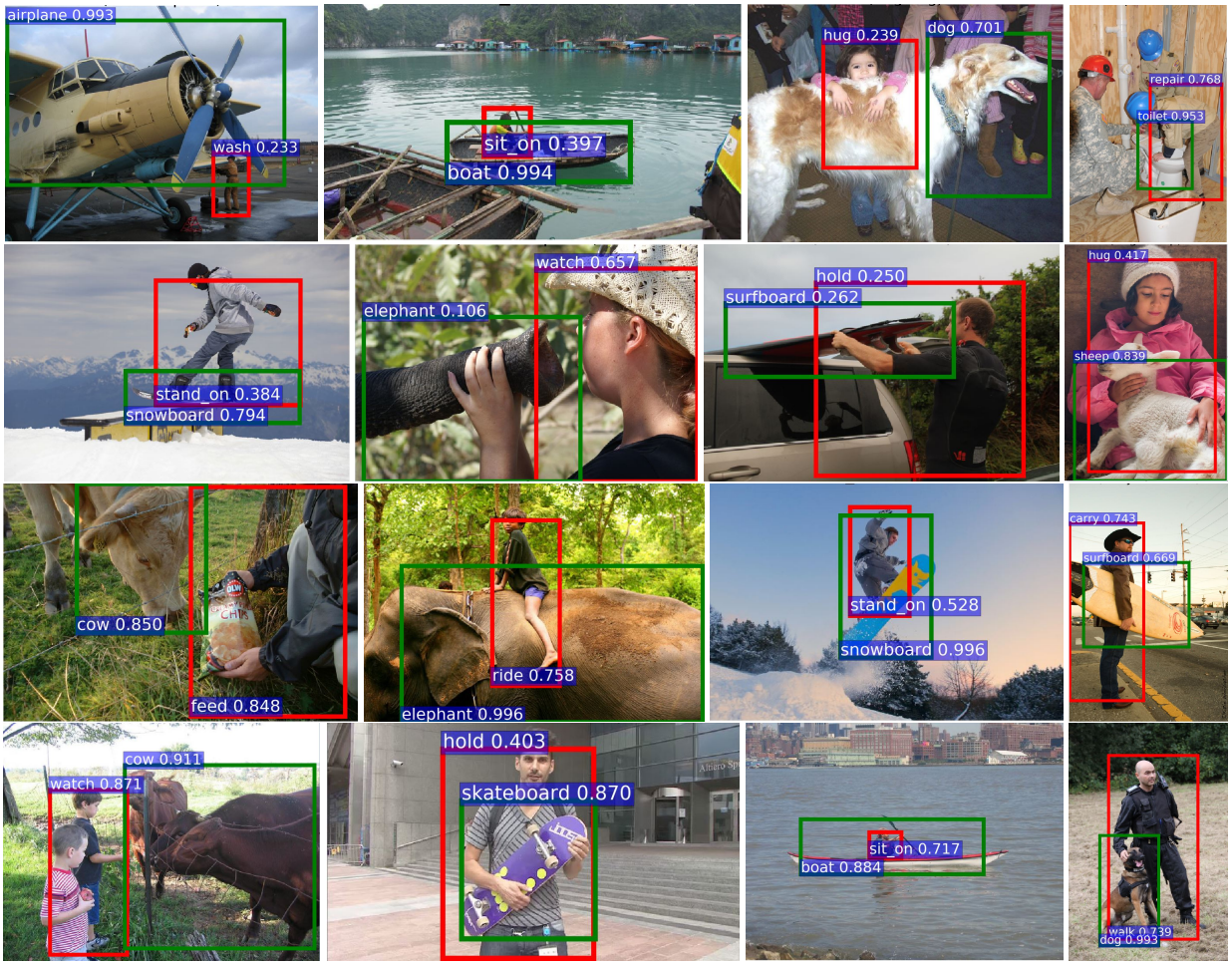

Figure 7: Qualitative results for zero shot learning experiments. The above HOI categories are unseen in training data. Our model is capable of composing novel combinations of verbs and objects as a result of our multi-task learning architecture. The model is able to learn the concepts of verbs such as hugging, holding, feeding, watching and sitting on, and transfers them to known objects and entities such as sheep, dogs, skateboards, boats, and airplanes.

object pairs which are absent from the HICO-DET labeled set as shown in Fig. 6, demonstrating example detections of new HOI categories beyond the entire labeled set. For example, we detect results of "watch horse", "herd elephant" and "aalk surfboard", respectively, while these classes are never annotated or included in the 600 HOI categories comprising the HICO-DET dataset. These examples highlight our model's potential for scalable human-object interaction recognition. More qualitative results for zero-shot detections are shown in the supplementary material.

5. Conclusion

In this work, we introduce a zero-shot learning approach towards scaling human-object interaction recognition to the long tail of categories. We present a model that factorizes HOI detection into disentangled verb and object networks with a shared early trunk, and train the model using a multi-task objective. We demonstrate that this model is able to both perform comparably to the state-of-the-art in fully-supervised HOI detection, while simultaneously achieving effective zero-shot detection of previously unseen verb-object pairs.

Future work includes extending the zero-shot learning approach to be able to more explicitly leverage stronger information on the structured spatial relationships in HOI, as well as to model temporal relationships in video data. Finally, we hope to extend this compositional framework beyond simple verb-object pairs to richer phrase and sentence queries.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *Proceedings*

of the IEEE Conference on Computer Vision and Pattern Recognition, pages 819–826, 2013.

[2] Z. Al-Halah and R. Stiefelhagen. How to transfer? zero-shot object recognition via hierarchical transfer of semantic attributes. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 837–843. IEEE, 2015.

[3] M. Bucher, S. Herbin, and F. Jurie. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In *European Conference on Computer Vision*, pages 730–746. Springer, 2016.

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050*, 2016.

[5] 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.

[6] V. Delaitre, J. Sivic, and I. Laptev. Learning person-object interactions for action recognition in still images. In *Advances in neural information processing systems*, pages 1503–1511, 2011.

[7] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In *European Conference on Computer Vision*, pages 158–172. Springer, 2012.

[8] Z. Fu, T. Xiang, E. Kodirov, and S. Gong. Zero-shot object recognition by semantic manifold distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2635–2644, 2015.

[9] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[10] R. Girshick. Fast r-cnn. In *International Conference on Computer Vision (ICCV)*, 2015.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition*, 2014.

[12] G. Gkioxari, R. B. Girshick, P. Dollár, and K. He. Detecting and recognizing human-object interactions. *CoRR*, abs/1704.07333, 2017.

[13] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2712–2719, 2013.

[14] S. Gupta and J. Malik. Visual semantic role labeling. *CoRR*, abs/1505.04474, 2015.

[15] J.-F. Hu, W.-S. Zheng, J. Lai, S. Gong, and T. Xiang. Recognising human-object interaction via exemplar based modelling. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3144–3151, 2013.

[16] D. Jayaraman and K. Grauman. Zero-shot recognition with unreliable attributes. In *Advances in Neural Information Processing Systems*, pages 3464–3472, 2014.

[17] D.-T. Le, J. Uijlings, and R. Bernardi. Tuhoi: Trento universal human object interaction dataset. *V&L Net 2014*, 2, 2014.

[18] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer, 2016.

[19] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3177–3184. IEEE, 2011.

[20] A. Mallya and S. Lazebnik. Learning models for actions and person-object interactions with transfer to question answering. In *European Conference on Computer Vision*, pages 414–428. Springer, 2016.

[21] T. Mensink, E. Gavves, and C. G. Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2441–2448, 2014.

[22] A. Prest. *Weakly supervised methods for learning actions and objects*. PhD thesis, ETH Zürich, 2012.

[23] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[24] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 568–576. Curran Associates, Inc., 2014.

[25] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 69–77, 2016.

[26] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 17–24. IEEE, 2010.

[27] X. L. H. Z. J. D. Yu-Wei Chao, Yunfan Liu. Learning to detect human-object interactions. 2017.

[28] Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6034–6042, 2016.