Discovering Human Interactions with Novel Objects via Zero-Shot Learning: Supplementary Material

A. Architecture

Here we report the details of our proposed model.

Backbone Our model is built upon the ResNet50 [4] with Feature Pyramid Network (FPN) [6]. Let $\{C_2, C_3, C_4, C_5\}$ denote the output of the last residual block of conv2, conv3, conv4, and conv5 at ResNet50. FPN calculates a feature pyramid $\{P_2, P_3, P_4, P_5, P_6\}$ based on $\{C_2, C_3, C_4, C_5\}$, where P_i is a feature map with 256 channels.

HO-RPN Given the feature pyramid $\{P_3, P_4, P_5, P_6\}$ from the backbone network, we use two separate branches to generate human and object region proposals. The human branch uses the original region proposal network (RPN) [10] with FPN. At each sliding window position, specifically, 3 anchor boxes with different aspect ratios $\{1{:}2,\ 1{:}1,\ 2{:}1\}$ are generated. At each level in the feature pyramid, we perform a 3×3 convolution on P_i to obtain a feature map H_i (with 256 channels), followed by two sibling 1×1 convolutions for classification and regression respectively. Here the classification is to predict the score of human anchor boxes and the regression is to adjust its shape. Over all levels in the feature pyramid, the top K ($K = 8$ in our experiments) human anchor boxes after the non-maximum suppression (NMS) are kept.

In the object branch, similar to RPN, the first step is to perform a 3×3 convolution on P_i to get a feature map O_i (with 256 channels). Here we still use a 1×1 convolution to regress the anchor boxes, while a relational network is used to predict its score. Specifically, at each sliding window position, the relational network takes as input its 256-D feature in O_i and the 256-D feature of top K human anchor boxes from $\{H_i\}$, and outputs the score of 3 anchor boxes. In Eq.(3), $g(\cdot) : \mathbb{R}^{512} \mapsto \mathbb{R}^{256}$ and $f(\cdot) : \mathbb{R}^{256} \mapsto \mathbb{R}^3$ are both multi-layer perceptrons (MLPs). Over all levels in the feature pyramid, the top 100 object region proposals after NMS are kept.

Head Network The head network is used to recognize the category of generated region proposals and further regress the box. We use RoIAlign [3] to extract the feature of the generated region proposals. To be compatible with novel object categories, we use a class-agnostic regressor and design a zero-shot classification module upon the softmax classifier of seen categories.

Zero-shot Classification Given a generated region proposal, the softmax classifier first predicts the probability of seen categories. If none of them have a score ≥ 0.1, we then estimate the probability of unseen categories using Eq.(5) and Eq.(6). When estimating the semantic embedding $\mathbf{e}$ in Eq.(5), the top 5 seen categories are used (please see Appendix D for more information). If the similarity of $\mathbf{e}$ to any unseen category is less than 0.2, it will be seen as a background, *i.e.*, $s_y = 0, \forall y \in \mathcal{Y}$.

Verb Prediction Given a human-object pair, we first extract the feature (1024-D) within the human bounding box, object bounding box, and their union region using RoIAlign (with 7×7 resolution). For each verb $v \in \mathcal{V}$, we use a simple MLP, $h_v(\cdot) : \mathbb{R}^{1024 \times 3} \mapsto \mathbb{R}$, to predict its probability.

B. Training Details

The various modules in our proposed model are trained jointly. The overall loss is the sum of (1) the binary classification loss and regression loss for the HO-RPN, (2) the cross-entropy loss and regression loss for the head network, (3) the binary classification loss for the verb prediction.

For training HO-RPN, we sample at most 64 anchor boxes per image. The ratio of foreground boxes to background boxes is set to 1:1. If the box has an Intersection-over-Union (IoU) ≥ 0.5 with the ground truth, it will be seen as the foreground. For training the head network, we sample at most 64 generated region proposals with a ratio of 1:1 of foreground boxes to background boxes. For training the verb prediction module, we sample at most 16 human-object pairs with a ratio of 1:1 of positives (*i.e.*, the human is interacting with the object) to negatives.

C. Seen/unseen Split

We simulate the zero-shot scenario on V-COCO [2] and HICO-DET [1] datasets by partitioning the 80 MS-COCO categories [7] into `seen` and `unseen` sets. Following previous works [8, 9], we construct the split based on the statistics of annotated HOI samples. Specifically, we sort

Supercategory	Category	#samples	cum.	Supercategory	Category	#samples	cum.
Accessory	handbad	288	0.05	Indoor	vase	84	0.02
	suitcase	687	0.16		hair drier	102	0.05
	tie	784	0.29		clock	222	0.11
	backpack	1738	0.57		teddy bear	304	0.19
	umbrella	2624	1.00		toothbrush	454	0.32
Animal	zebra	43	0.01		scissors	469	0.44
	bear	89	0.02		book	2063	1.00
	bird	313	0.06	Kitchen	bowl	135	0.03
	cat	409	0.11		fork	295	0.09
	cow	414	0.17		spoon	372	0.17
	giraffe	420	0.22		wine glass	632	0.31
	sheep	494	0.29		cup	797	0.48
	elephant	862	0.40		bottle	968	0.69
	dog	1145	0.55		knife	1457	1.00
	horse	3404	1.00	Outdoor	stop sign	57	0.03
Appliance	toaster	7	0.02		fire hydrant	139	0.10
	microwave	28	0.10		traffic light	143	0.18
	sink	71	0.29		parking meter	144	0.25
	refrigerator	91	0.55		bench	1448	1.00
	oven	164	1.00	Sports	baseball glove	348	0.02
Electronic	mouse	110	0.03		tennis racket	661	0.05
	tv	219	0.09		surfboard	1350	0.13
	remote	286	0.17		frisbee	1491	0.20
	keyboard	340	0.26		kite	1709	0.30
	laptop	1285	0.60		baseball bat	1768	0.39
	cell phone	1468	1.00		sports ball	1878	0.49
Food	orange	60	0.01		skis	1933	0.60
	broccoli	100	0.04		snowboard	2334	0.72
	carrot	313	0.10		skateboard	5262	1.00
	apple	354	0.18	Vehicle	truck	620	0.03
	sandwich	418	0.27		train	670	0.07
	hot dog	536	0.39		airplane	950	0.12
	dount	537	0.51		car	1261	0.19
	pizza	591	0.64		bus	1651	0.28
	banana	694	0.79		boat	3285	0.46
	cake	959	1.00		motorcycle	4699	0.72
Furniture	potted plant	73	0.02		bicycle	5202	1.00
	toilet	179	0.06				
	bed	595	0.19				
	couch	859	0.38				
	chair	1293	0.67				
	dining table	1476	1.00				

Table 1: **Our seen/unseen split**. We sort classes per supercategory in ascending order and calculate the cumulative percentage (cum.). The rare 20% classes (highlighted by gray shade) are selected as unseen.

classes per supercategory in ascending order based on the number of samples in V-COCO `train` set and HICO-DET `train` set. For each supercategory, we calculate the cumulative percentage and select the 20% rare classes as unseen classes. Table 1 reports our `seen`/`unseen` split.

D. Discussion on Zero-shot Classification

In Table 2, we study how many seen categories should be considered in Eq.(5) when we estimate the semantic embedding vector $\mathbf{e}$. As $\mathbf{e}$ is used to detect novel objects, we use the AP@IoU=0.5 over unseen object categories to investigate its effects. The experiments are conducted on V-COCO `val` set. As shown, when only one seen category is used, our method can only achieve an AP of 7.18. The result increases to 12.49 when the top 3 seen categories are used. At $K = 5$, our method achieves the best performance. As more seen categories are considered (*e.g.*, $K = 7$), the performance slightly drops since unrelated categories may be included. Hence, we choose $K = 5$ for the rest of the experiments.

In Table 3, for each unseen category, we report the top 3 frequently used seen categories and their averaged weights (after normalization) when estimating its semantic embedding $\mathbf{e}$ in Eq.(5). We observe that the unseen categories are generally expressed by its semantically or functionally similar object categories. For example, the unseen category "bed" is expressed as the weighted sum of "couch", "chair", and "dining table"; "cow" is expressed as the weighted sum of "horse", "sheep" and "dog"; "surfboard" is expressed as the weighted sum of "snowboard", "boat", and "skateboard". For our method, having related seen categories is a key factor for the success of detecting novel objects.

K	1	2	3	5	7
AP@IoU=0.5	7.18	11.98	12.49	12.55	12.53

Table 2: Experimental results of using various number of seen categories in Eq.(5).

Unseen category	Top 3 frequently used seen categories			AP
airplane	boat (0.367)	person (0.217)	bench (0.156)	0.6
apple	banana (0.453)	donut (0.188)	cake (0.071)	26.1
baseball glove	sports ball (0.475)	baseball bat (0.297)	chair (0.026)	4.7
bed	couch (0.328)	chair (0.310)	dining table (0.129)	43.4
bird	dog (0.329)	baseball bat (0.162)	horse (0.132)	3.5
bowl	sports ball (0.439)	baseball bat (0.236)	cup (0.111)	2.0
broccoli	pizza (0.421)	sandwich (0.266)	hot dog (0.079)	0.4
car	bench (0.123)	chair (0.113)	dining table (0.076)	11.8
carrot	umbrella (0.738)	banana (0.086)	cake (0.069)	0.0
cat	dog (0.539)	couch (0.068)	hot dog (0.060)	1.1
cow	horse (0.723)	sheep (0.145)	dog (0.026)	12.6
fire hydrant	parking meter (0.894)	skateboard (0.036)	dog (0.013)	100.0
fork	dining table (0.554)	banana (0.186)	umbrella (0.117)	3.6
frisbee	sports ball (0.357)	kite (0.171)	umbrella (0.080)	26.6
handbag	laptop (0.453)	book (0.141)	cell phone (0.083)	6.5
microwave	refrigerator (0.396)	oven (0.323)	dining table (0.068)	-
mouse	keyboard (0.559)	laptop (0.245)	cell phone (0.076)	1.0
orange	umbrella (0.599)	tie (0.152)	banana (0.151)	0.5
potted plant	tie (0.487)	book (0.375)	cake (0.077)	-
spoon	knife (0.267)	cup (0.150)	cake (0.137)	1.9
suitcase	backpack (0.702)	chair (0.073)	bench (0.027)	10.7
surfboard	snowboard (0.312)	boat (0.213)	skateboard (0.205)	31.1
teddy bear	umbrella (0.247)	elephant (0.212)	book (0.108)	12.1
toaster	book (0.188)	laptop (0.159)	refrigerator (0.149)	-
toilet	bench (0.417)	cell phone (0.163)	sink (0.091)	0.2
train	bus (0.785)	bench (0.062)	boat (0.039)	25.9
truck	bus (0.348)	boat (0.335)	umbrella (0.057)	6.1
vase	wine glass (0.268)	book (0.228)	dining table (0.146)	-
zebra	elephant (0.822)	giraffe (0.044)	horse (0.041)	-

Table 3: **The top 3 frequently used seen categories** for each unseen category. The results are evaluated on V-COCO `val` set. Here we do not show the category "bear", "hair drier", *etc*., since no prediction is made by our model for those unseen categories on V-COCO `val` set.

E. Human-Novel-Object Interactions

In Table 4, we show more details of our experiments on human-novel-object interaction detection. Based on our `seen`/`unseen` split, there are 199 human-novel-object interactions on the HICO-DET dataset. We observe that 74 out of them are missing from our detections (*i.e*., an AP of 0.0). In Table 4, we only show the first 20 missing interactions (in alphabetical order). There are two possible reasons for the missing detection. First, the verb prediction is overfitted to a specific object category. For instance, the verb "blocking" has 60 training samples, while all the interacting objects are "sports ball". In this case, our verb

Best 20 Interactions	AP	Worst 20 Interactions	AP
feeding zebra	29.35	blocking frisbee	0.00
watching zebra	25.34	buying apple	0.00
lying on bed	20.69	buying orange	0.00
carrying surfboard	17.73	carrying carrot	0.00
operating microwave	17.39	carrying potted plant	0.00
washing car	17.17	carrying teddy bear	0.00
dragging suitcase	16.75	checking clock	0.00
feeding cow	15.18	controling tv	0.00
flying airplane	14.69	cooking carrot	0.00
holding broccoli	14.56	cutting carrot	0.00
holding surfboard	14.04	cutting orange	0.00
directing airplane	12.30	eating carrot	0.00
carrying suitcase	11.57	eating orange	0.00
catching frisbee	10.19	holding bowl	0.00
eating broccoli	10.07	holding carrot	0.00
eating apple	9.72	holding clock	0.00
carrying handbag	9.52	holding fork	0.00
throwing frisbee	9.09	holding hair drier	0.00
standing on surfboard	8.76	holding orange	0.00
wearing baseball glove	7.71	holding potted plant	0.00

Table 4: **Human-novel-object detection.** The best and worst 20 human-novel-object interactions detected by our model.

prediction module is overfitted to the interaction "blocking sports ball", resulting in the missing detection of "blocking frisbee". In comparison, the verb "feeding" have 5 different object categories in the training set, *i.e.*, "dog", "horse", "sheep", "elephant" and "giraffe". In this case, our model suffers less from the overfitting issue and can generalize well to the interaction "feeding zebra". Second, we observe that it is difficult for our model to detect some unseen categories, *e.g.*, "carrot", "bowl", "orange", *etc*. On the one hand, it is because they are often small-scale objects. One the other hand, it is because no closely related seen categories exist based on our `seen/unseen` split. For instance, as shown in Table 3, "carrot" is expressed as the weighted sum of dissimilar categories "umbrella". "banana" and "cake".

F. Visual Genome (VG) Test Set

Apart from simulating the zero-shot scenario by partitioning the 80 MS-COCO object categories, we also construct a test set from Visual Genome (VG) [5] with additional 30 novel object categories (see Table 5). Notice that the words used to describe the human-object interactions in the VG dataset are very different from the HICO-DET dataset (which is used for training our model). For instance, the annotations of the first image in Figure 1 are "reading screen", "playing piano", "sitting on chair", "seated at piano", *etc*. However, "screen", "playing", "seated" are not defined in HICO-DET dataset. Due to this mismatch, it is difficult to provide the quantitive results on our collected VG test set. For this reason, we mainly use qualitative evaluation. Figure 1 and Figure 2 show more qualitative results of our proposed model. The interactions are depicted if its score is greater than 0.5 for known objects and 0.2 for novel objects.

No.	Category name	#images	No.	Category name	#images
1	barrel	6	16	jetski	5
2	blender	13	17	lemon	5
3	box	30	18	microphone	32
4	bread	15	19	monitor	2
5	broom	5	20	newspaper	32
6	brush	9	21	paddle	46
7	camera	85	22	pail	4
8	carriage	30	23	pen	20
9	dolphin	8	24	piano	12
10	drum	4	25	pushcart	7
11	fishing pole	10	26	rose	5
12	flowers	11	27	tomato	5
13	guitar	23	28	tractor	4
14	gun	10	29	turkey	3
15	ipad	4	30	wheelchair	9

Table 5: **Visual Genome test set.**

References

[1] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 1

[2] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 1

[3] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[5] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 3

[6] T. Lin, P. Dollr, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*. 2014. 1

[8] Shafin Rahman, Salman Khan, and Nick Barnes. Polarity loss for zero-shot object detection. *arXiv preprint arXiv:1811.08982*, 2018. 1

[9] Shafin Rahman, Salman Khan, and Nick Barnes. Transductive learning for zero-shot object detection. *International Conference on Computer Vision (ICCV)*, 2019. 1

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1

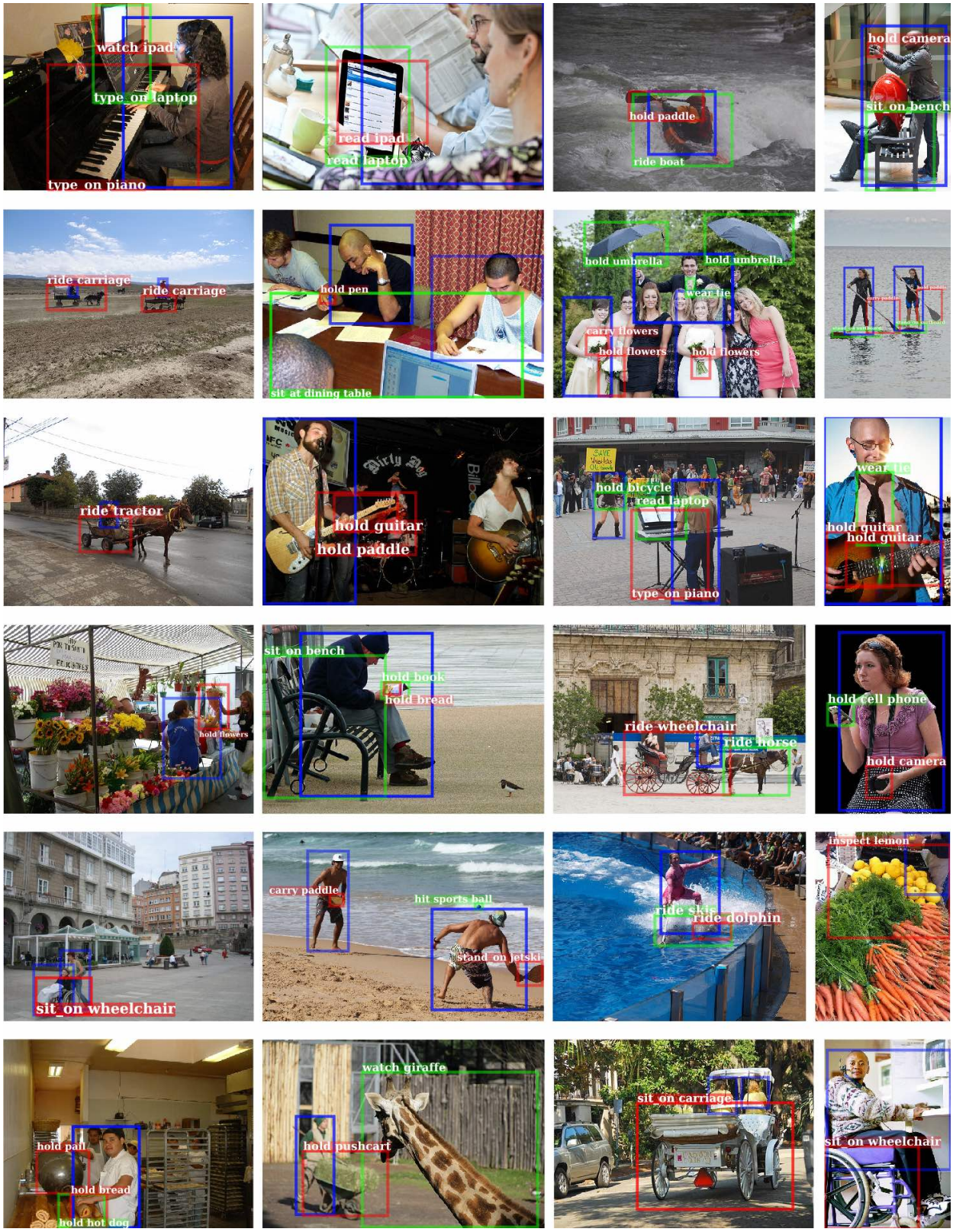

Figure 1: Qualitative results of our model on human-novel-object interaction detection. Seen object categories are highlighted by green, while novel object categories are highlighted by red.

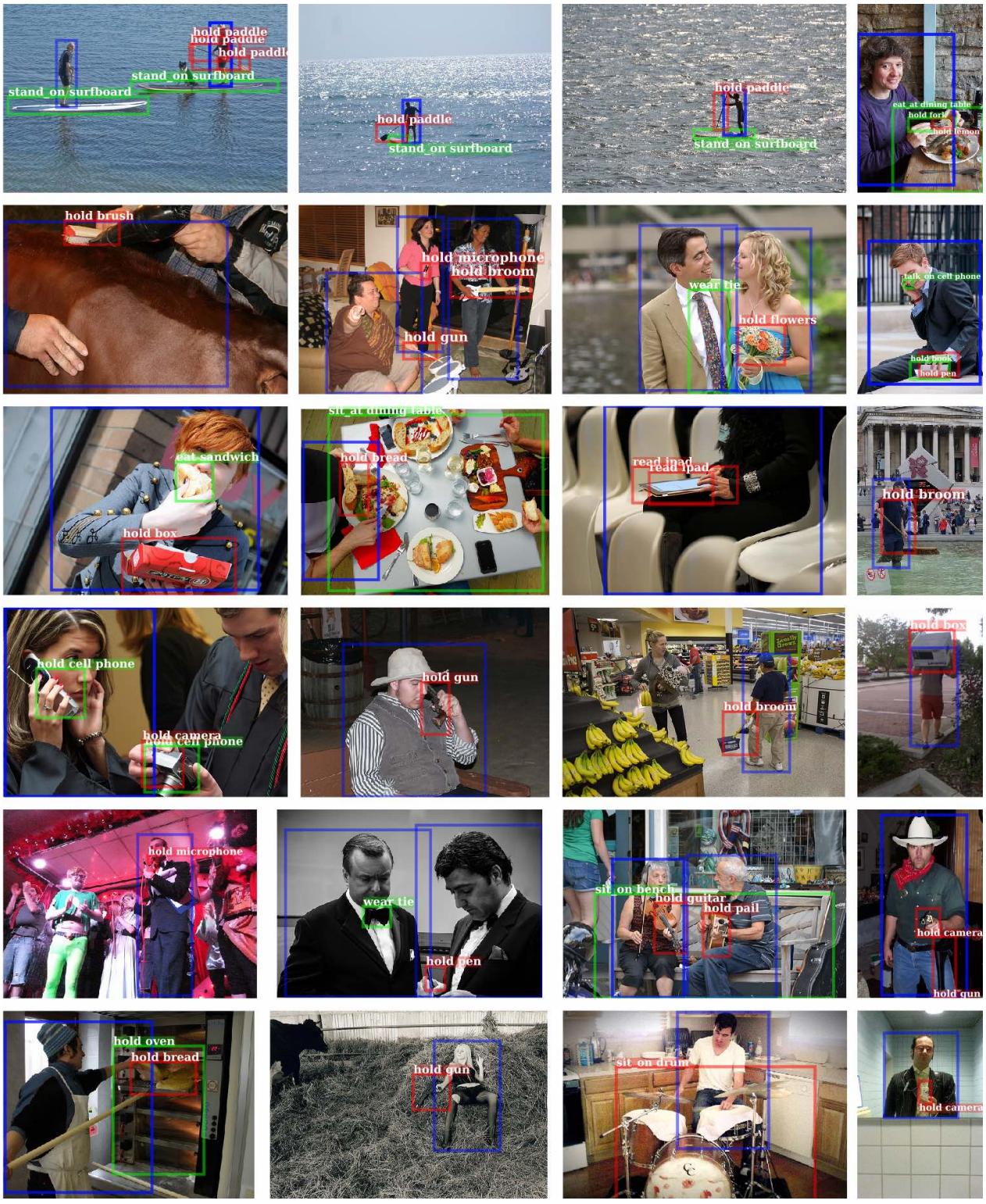

Figure 2: Qualitative results of our model on human-novel-object interaction detection. Seen object categories are highlighted by green, while novel object categories are highlighted by red.