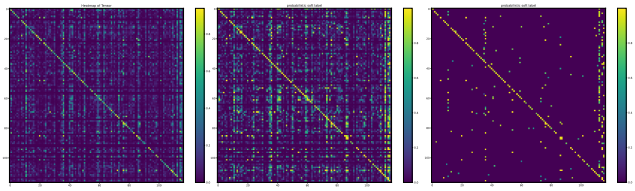
Exploring Pose-Aware Human-Object Interaction via Hybrid Learning

Supplementary Material



(a) baseline (b) Ours (c) Ours-top3

Figure 1. This figure illustrates the average soft labels generated on the HICO-DET dataset, representing correlations between interactions. Each map's size is 117×117, where the i^{th} row signifies the correlation coefficients between the i^{th} interaction category and the remaining 117 categories.

1. Implicit Relationships between Interactions

The hybrid learning method effectively captures implicit connections among human-object interactions. In this section, we offer more intuitive experiments to substantiate this assertion. We visualized the probabilistic soft labels for all 117 actions, depicting implicit relationships among HOI interactions. Specifically, we conduct human-object interaction detection on the entire training set of 38,118 images from HICO-DET. We obtain 117-dimensional predictions for the most confident interaction pairs within each image, representing the correlation of each action with other interactions. Finally, we accumulate the average predictions for the 117 action classes in the dataset, resulting in a 117×117 correlation matrix. As depicted in Figure 1, (a) illustrates the average correlation matrix of the baseline model, while (b) showcases our correlation coefficient matrix. From the most intuitive perspective, our correlation coefficient map is brighter both on the diagonal and in certain areas. This indicates a more comprehensive consideration of potential interaction in our approach, as opposed to treating other interaction pairs merely as negative samples. In Figure 1(c), we present the filtered correlation coefficient map produced by our method. Each row retains the top three highest confidence correlation coefficients to offer a more intuitive insight into implicit relationships. We notice relatively higher coefficients on average in the 113^{th} and 115^{th} columns of the map, corresponding to *'watch'* and *'wear'* interactions, respectively. This aligns with our expectations as *'wear'* and *'watch'* actions commonly exist in interactions between human and objects.

To further delve into the implicit correlations depicted in hybrid learning, We list representative actions along with highly correlated interactions in Table 1. We categorize interactions between humans and objects into four primary groups: spatial information related (S), human pose related (P), language semantic related (L), and object related (O). For instance, in the first row of the table 1, the *'ride'* action is highly correlated with spatial and pose features, denoted as (S,P). Among the top 5 interactions most correlated with *'ride'* in the soft labels, *'sit at'* and *'sit on'* exhibit resemblances primarily in spatial characteristics, whereas "*'straddle'* demonstrates similarities in both spatial and pose features. This signifies the efficacy of our method in delineating implicit relationships between interactions through the generated soft labels. Similarly, for hand pose related action like *'swing'*, foot related action *'kick'*, object and facial pose related action *'eat'*, the soft labels effectively capture potential labels associated with them. It's worth noting that the introduction of soft labels provides additional supervision for rare classes lacking training samples. For instance, in the soft labels for the *'kick'* action, the generated *'dribble'* soft label supplements its supervision with features extracted from the *'kick'* action within the feature space. On the other hand, when examining the *'no interaction'* class, our hybrid learning method didn't significantly increase the False Positive Rate (FPR) of genuine negative samples.

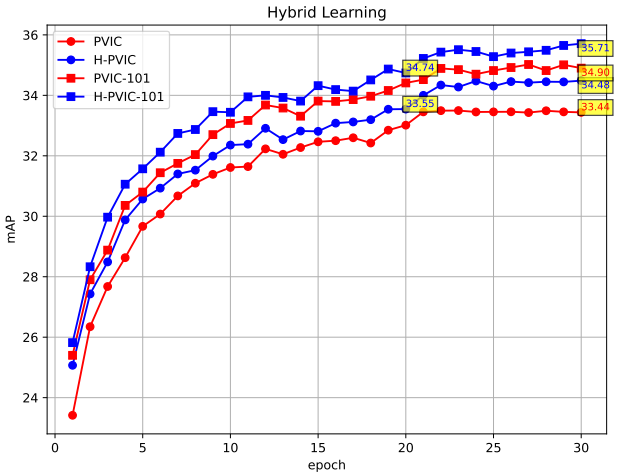


Figure 2. Illustrating the HOI detection performance of the baseline model PViC[**?**] and $\mathcal{H}$-PViC throughout the entire training process. $\mathcal{H}$-PViC refers to the baseline model PViC with the hybrid learning. The red line represents $\mathcal{H}$-PViC, and the blue line represents PViC. Circle and square markers correspond to ResNet50 and ResNet101 respectively.

2. Hybrid Learning's Superiority

To further demonstrate the effectiveness and efficiency of hybrid learning, we showcase the HOI detection perfor-

Backbone	Performed Action	Top-5 Correlated Interactions in Probabilistic Soft Labels				
R50	*ride* (S,P)	ride: 0.84	wear: 0.56	sit-on: 0.50	straddle: 0.50	sit-at: 0.42
Swin-L		ride: 0.99	wear: 0.95	sit-at: 0.85	straddle: 0.72	sit-on: 0.62
R50	*swing* (P)	swing: 0.80	operate: 0.69	wield: 0.69	watch: 0.68	hold: 0.68
Swin-L		swing: 0.98	wield: 0.92	hold: 0.90	operate: 0.86	wear: 0.84
R50	*eat* (P,O)	eat: 0.84	hold: 0.68	talk on: 0.48	drink with: 0.46	brush with: 0.39
Swin-L		eat: 0.98	hold: 0.86	wear: 0.85	brush with: 0.79	drink with: 0.57
R50	*kick* (P)	kick: 0.81	ride: 0.51	stand on: 0.43	dribble: 0.31	lasso: 0.31
Swin-L		kick: 0.96	wear: 0.87	stand on: 0.81	ride: 0.67	dribble: 0.37
R50	*load* (L)	load: 0.58	check: 0.64	pay: 0.45	install: 0.39	hunt: 0.34
Swin-L		load: 0.96	install: 0.77	watch: 0.73	move: 0.62	check: 0.60
R50	*no interaction*	no: 0.65	watch: 0.13	check: 0.07	hit: 0.07	lift: 0.06
Swin-L		no: 0.97	watch: 0.33	hit: 0.13	serve: 0.08	read: 0.07

Table 1. This table represents the probabilistic soft labels generated for six different actions. It showcases the top five interactions with the highest correlation coefficients to the performed action, along with their corresponding confidence scores. The values of the probabilistic soft labels are derived from the averaged soft labels generated across all images in the HICO-DET training set.

Backbone	Method	Train Time	inf. GFLOPs
ResNet50	PViC	6h 13min	73.1G
	$\mathcal{H}$-PViC	6h 40min	73.1G
ResNet101	PViC	7h 03min	132.8G
	$\mathcal{H}$-PViC	7h 42min	132.8G

Table 2. Comparisons of training and inference efficiency between employing hybrid learning and the original method. The experiments were conducted on 8 GTX 1080Ti GPUs, training for 30 epochs. The 'inf.' in the table stands for 'inference'.

mance of $\mathcal{H}$-PViC (model with hybrid learning) compared to PViC throughout each epoch of the training process in Figure 2. It's worth noting that due to no modifications in the model architecture, the initial performance of both $\mathcal{H}$-PViC and PViC models remains identical: 13.68 mAP for ResNet50, 14.1 mAP for ResNet101, and 19.26 mAP for Swin-Large. As depicted in the line plot, after training with supervision from probabilistic soft labels, $\mathcal{H}$-PViC consistently outperformed baseline across the 30 epochs of training. Notably, this advantage did not diminish in the later stages of training. This underscores that the introduction of hybrid learning not only facilitates smoother gradient descent but also enables the model to converge towards a more favorable local optimum. Quantitative analysis indicates that, under identical optimizer and learning rate settings, $\mathcal{H}$-PViC achieves the performance of a fully trained PViC in only 20 epochs. Additionally, fully training with soft label supervision results in a 1.0 mAP improvement. This demonstrate that hybrid learning can expedite convergence and enhance the model's generalization capabilities without architectural changes or additional training data.

On the other hand, we analyze the efficiency of hybrid learning, as depicted in Table 2. After implementing the probabilistic soft label supervision, the training time increase slightly, specifically by 7.2% for ResNet50 and 9.2% for ResNet101. Additionally, without altering the model architecture, $\mathcal{H}$-PViC maintains an equivalent inference GFLOPS as the original PViC.

3. Distillation Hyper-parameter Optimization

In Hybrid Learning, we generated probabilistic soft labels by fully-trained interaction head to mitigate challenges associated with mislabeling and sparse annotations. While soft labels offer a more refined form of supervision, they may also introduce noise due to potential misinterpretations within the interaction head. In the decay scheme of hybrid learning, we devised a stepwise reduction of the hyperparameter λ to gradually diminish the influence of soft labels. This strategic adjustment empowers the model to transcend the cognitive constraints of the interaction head, mitigating its biases in the later phases of the training process. However, we can obtain a simpler approach by distilling the generated soft labels. Specifically, we set a threshold t_d where any soft label with a confidence below t_d is set to zero. This filtering process aims to eliminate potentially erroneous soft labels generated by the interaction head. We conducted experiments by setting the threshold from 0.0 to 1.0 in increments of 0.1, as shown in Figure 3.

Specifically, when $t_d = 1.0$, it's equivalent to not uti-

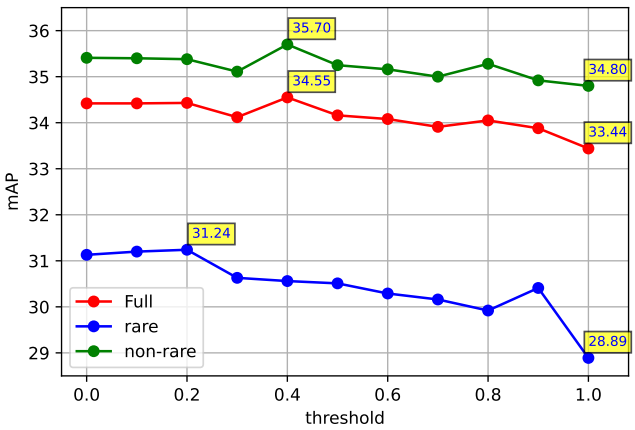


Figure 3. The HOI detection performance is assessed across subsets—'Full,' 'Rare,' and 'Non-rare'—based on varying distillation hyper-parameters t_d. The line plot utilizes red, blue, and green colors to represent the performance on 'Full,' 'Rare,' and 'Non-rare' subsets respectively.

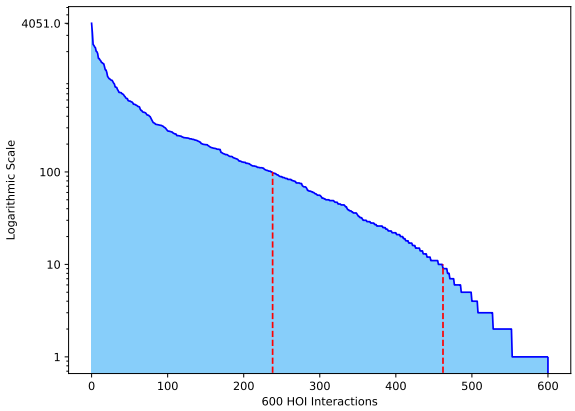


Figure 4. Illustrating the frequency of occurrence for HOI interaction triplets in the HICO-DET training dataset. Due to the significant variance in the occurrences of HOI triplets, we use a logarithmic scale. The dataset exhibits a clear long-tail distribution.

Samples num	HOI Interaction	mAP	mRec
5	hug cow	96.97	100.00
6	kiss cat	96.59	100.00
2	hold zebra	95.45	100.00
9	paint fire hydrant	92.06	100.00
4	pet zebra	90.91	100.00
6	wash bus	90.00	100.00
8	wash train	87.88	100.00
2	clean bed	87.88	100.00
1	sign baseball bat	87.50	100.00
3	kiss teddy bear	86.46	100.00
8	feed zebra	85.91	100.00
1	wash toothbrush	84.75	100.00
3	eat orange	84.09	100.00
1	clean microwave	83.64	100.00
3	wash motorcycle	83.33	100.00

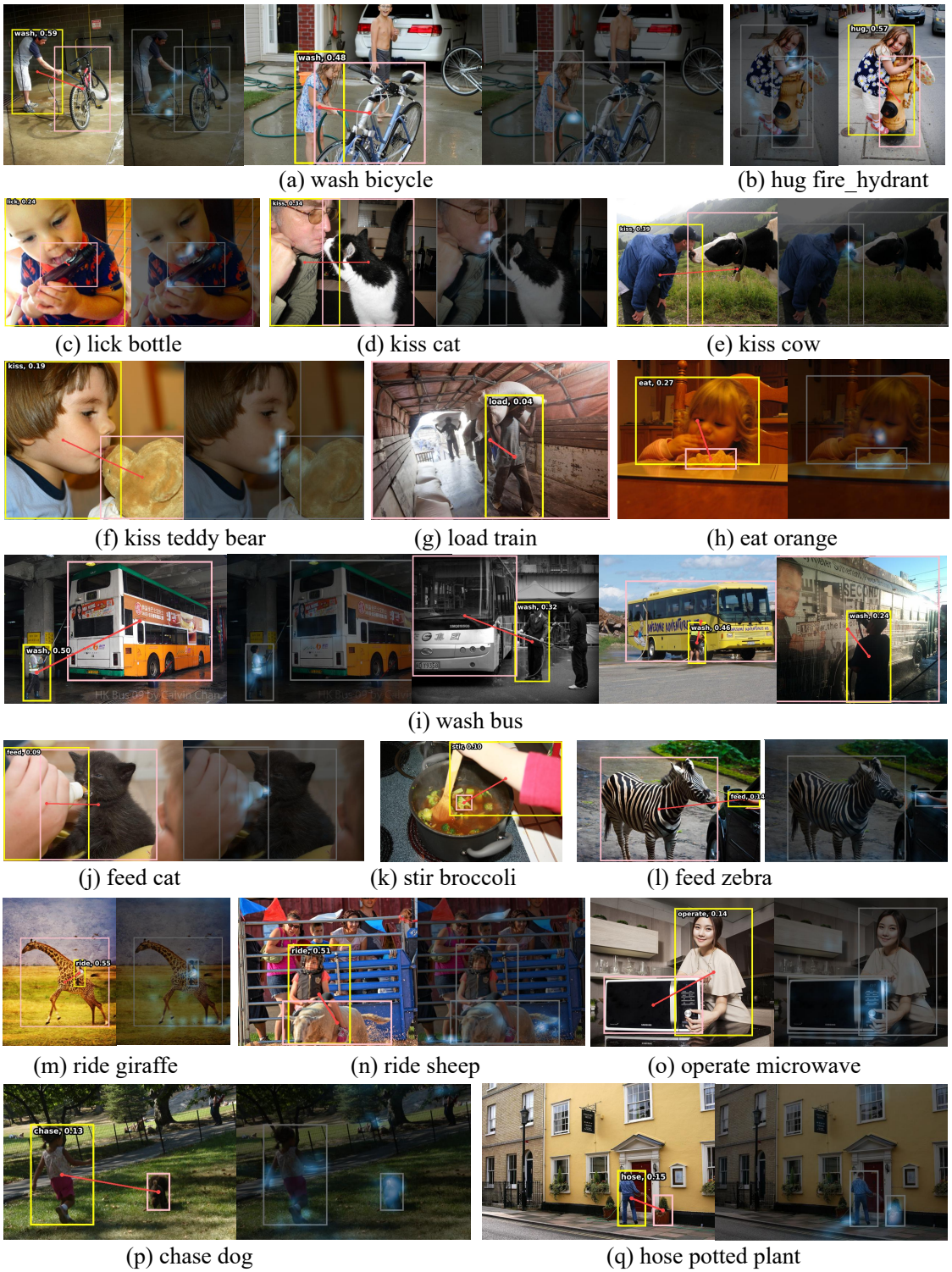
Table 3. We present our detector's detection performance on rare classes, listing the top 15 interactions.

lizing the soft label supervision, whereas $t_d = 0$ means no label filtering being applied. Figure 3 illustrates that soft lable supervision is most effective when t_d is set to 0.4, achieving a 34.55 mAP on full classes detection. The performance trend of the detector on rare classes implies that as the threshold t_d decreases, allowing for the retention of more low-confidence labels, the average detection performance in identifying rare interaction tends to increase. This could be attributed to the relatively lower confidence of rare classes within the soft labels, hence reducing the threshold aids in retaining the implicit associations of rare classes by the interaction head. Overall, within the range of 0.2 to 0.6, the hyperparameter t_d appears to be more suitable for denoising. Training the model with distilled soft label supervision consistently improves its HOI detection performance, especially on rare classes.

4. Qualitative Results of Rare Interactions

In the earlier discussion, we emphasized the efficacy of hybrid learning in improving detection performance for rare classes. In this section, we offer a more detailed analysis and visualization of these results. First, we conduct a statistical analysis on the distribution of HOI categories within the HICO-DET dataset, as shown in Figure 4, revealing a clear long-tail distribution. In more detail, with the most frequent *'ride boat'* occurring 4051 times, while nearly 60% of HOI categories have fewer than 100 training samples, and 25% have fewer than 10 training samples. To address the aforemention issues, we implement hybrid learning, incorporating probabilistic soft label supervision into training process.

Our approach achieved an unprecedented 46.74 mAP on rare classes. We showcase the detection performance on some rare HOI interactions in Table 3. The previous method overly focused on object features when identifying interactions, resulting in overfitting to the objects involved. In contrast, our approach emphasizes exploring implicit relationships between predicates, accurately capturing the key elements in identifying interactions. For instance, interactions involving cows like *'herd cow'* have 245 samples, *'walk cow'* has 127, while *'hug cow'* only has 5 samples. Overfitting to objects might lead to misclassifying *'hug'* as *'walk'* or *'herd'*. Our approach emphasizes learning the potential connections between predicates, prioritizing human actions over objects. This led to achieving 96.97% mAP and 100% mRec for the *'hug cow'* interactions.



(a) wash bicycle (b) hug fire_hydrant

(c) lick bottle (d) kiss cat (e) kiss cow

(f) kiss teddy bear (g) load train (h) eat orange

(i) wash bus

(j) feed cat (k) stir broccoli (l) feed zebra

(m) ride giraffe (n) ride sheep (o) operate microwave

(p) chase dog (q) hose potted plant

Figure 5. Qualitative results of rare interactions. All interactions displayed here have no more than 10 training samples.