Human-Object Interaction Detection *without* Alignment Supervision

Mert Kilickaya,
kilickayamert@gmail.com
Arnold W.M. Smeulders
arnoldsmeulders@uva.nl

Visual Sensing Lab
University of Amsterdam
Amsterdam, Netherlands

Abstract

The goal of this paper is Human-object Interaction (HO-I) detection. HO-I detection aims to find interacting human-objects regions and classify their interaction from an image. Researchers obtain significant improvement in recent years by relying on strong HO-I alignment supervision from [5]. HO-I alignment supervision pairs humans with their interacted objects, and then aligns human-object pair(s) with their interaction categories. Since collecting such annotation is expensive, in this paper, we propose to detect HO-I without alignment supervision. We instead rely on image-level supervision that only enumerates existing interactions within the image without pointing where they happen. Our paper makes three contributions: *i)* We propose Align-Former, a visual-transformer based CNN that can detect HO-I with only image-level supervision. *ii)* Align-Former is equipped with HO-I align layer, that can learn to select appropriate targets to allow detector supervision. *iii)* We evaluate Align-Former on HICO-DET [5] and V-COCO [13], and show that Align-Former outperforms existing image-level supervised HO-I detectors by a large margin (**4.71**% mAP improvement from 16.14% $\rightarrow$ 20.85% on HICO-DET [5]).

1 Introduction

This paper strives for Human-object Interaction (HO-I) detection from an image. HO-I detection receives an astounding attention from the community recently [5, 6, 9, 10, 12, 14, 16, 18, 19, 20, 22, 24, 28], thanks to the large-scale benchmark of HICO-DET [5]. The goal is to identify the tuples of `<human, object, verb, noun>` from the input, where human-object is an interacting bounding box pair, and verb-noun is the interaction type, such as ride-horse.

To tackle this problem, researchers leverage strong HO-I alignment supervision, see Figure 1-(a). Annotators first draw a bounding box around all humans and objects, then align humans with the object-of-interaction (*e.g.*, rider and horse). Finally, they align the interaction category with each human-object pairs.

However, collecting such annotation is costly [1]. Annotation costs time, since in a typical image there are tens of potential human-object interactors, if not hundreds. One can instead rely on image-level HO-I annotations, see Figure 1-(b). Image-level annotations enumerate existing HO-I within the image, without specifying where they occur. Image-level annotations are much faster and cheaper to collect.

[1] Try-it-yourself! HICO-DET-Annotator

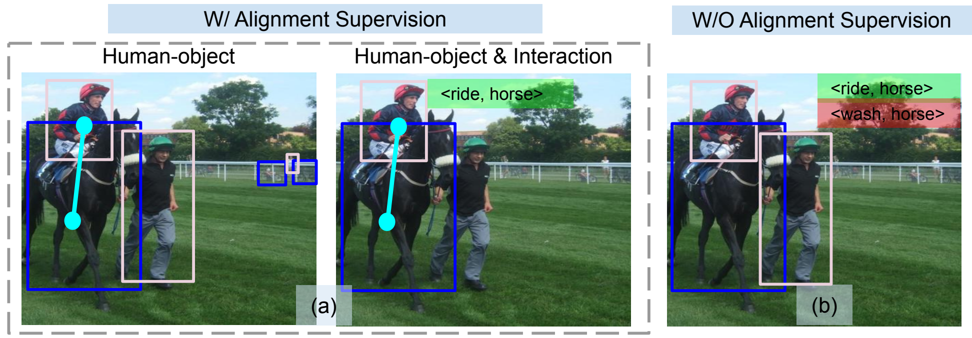

Figure 1: Alignment (left) *vs.* Image-level HO-I supervision (right). *a)* Alignment supervision annotates each human-objects, aligns humans to their interacting objects, then aligns human-objects to their type of interaction. *b)* Image-level supervision only lists existing interactions without pointing where they happen. Our goal is to detect HO-I *without* costly alignment supervision, by only using image-level labels.

There are few attempts to perform HO-I detection via image-level supervision [21, 31]. Initially, Zhang *et al.* [31] proposes a two-stream architecture based on Region-FCN [8], focusing on the regional appearance of subject-objects and spatial relations. Later, Kumaraswamy *et al.* [21] adapted this technique for HO-I detection, and improve it via an additional stream of human pose. These techniques yield remarkable results on HICO-DET benchmark [5] in the absence of alignment supervision. However, they are limited in three major ways: *i)* These methods isolate human-objects from their context via Region-of-Interest (RoI) pooling [11, 27], however, contextual information is crucial in understanding the interaction, *ii)* The authors propose multiple streams of context to circumvent the missing contextual information, which increases model complexity. Increased model complexity results in low performance on especially rarely represented HO-I (*i.e.*, `<ride, cow>`) as we will show. *iii)* Hand-crafted context (*i.e.*, body-pose configuration using key-points) may not be sufficient to account for the complexity of HO-I detection problem.

To that end, in this paper, we propose Align-Former, a visual-transformer-based architecture based on [4]. Align-Former is a single-stream HO-I detector that is trained end-to-end using image-level supervision only. Align-Former is equipped with a novel HO-I Align layer that learns to align a few candidate target HO-I with predictions, allowing detector supervision. The decision of alignment is based on geometric and visual priors that are crucial in HO-I detection.

This paper makes the following contributions:

I. We propose Align-Former, an end-to-end HO-I detector that is supervised via image-level annotation.

II. We equip Align-Former with a novel HO-I align layer, that learns to match few HO-I predictions with HO-I target(s), therefore allowing detector supervision.

III. We evaluate Align-Former on HICO-DET [5] and V-COCO [13], and show that Align-Former outperforms competing baselines with the same level of supervision (by **4.71** mAP) on the large-scale benchmark of HICO-DET [5], especially within the low-data regime of rare categories (by **6.17** mAP).

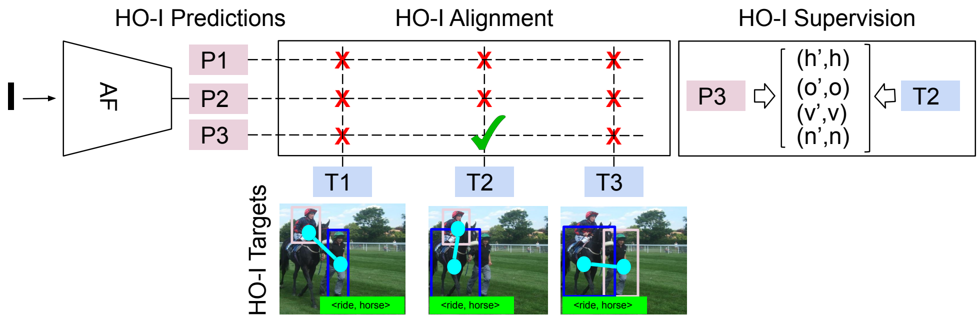

Figure 2: To perform HO-I detection via image-level supervision: *i)* Align-Former maps the input image I to HO-I predictions P . *ii)* We also prepare a set of HO-I targets by exhaustively matching human-object detections and list of interactions. *iii)* Finally, we find the least costly prediction-target pair(s) (*i.e.*, (T_2, P_3)) which will be used for detector supervision.

2 Related Work

Alignment-Supervised HO-I Detection. In HO-I detection, the goal is to find quadruplets of `<human,object,verb, noun>` where human-object are bounding boxes and verb-noun are interaction pairs like `<ride, horse>`. Initially, HICO-DET authors collect more than $150k$ instance annotations to match humans to their interacted object, as well as to their interaction categories. Then, there has been a surge in detecting HO-I, initially via two-stage techniques [5, 9, 12, 14, 16, 24], and later by one-stage architectures [6, 10, 20, 22, 28] leveraging costly strong alignment supervision, see Figure 1-(a).

In this work, our goal is to train HO-I detectors without alignment supervision, by only relying on image-level HO-I annotations.

HO-I Detection via Image-level Supervision. Few works attempt to train HO-I detectors by only image-level supervision [21, 31]. Initially, Zhang *et al.* [31] proposes a two-stream architecture based on Region-FCN [8] to model the subject-object region appearance and spatial relations. Later, Kumaraswamy *et al.* [21] extends this approach via additional pose-stream. These methods operate on the isolated appearance of human-objects, neglecting the crucial context. Consequently, they supplement Region-FCN with additional streams, increasing the model size, decreasing the performance.

To circumvent this, in this work, we propose a single-stream HO-I detector based on visual-transformer [4]. Our network naturally encodes the surrounding context of human-objects thanks to self-attention [29] and learns to align few candidate HO-I targets with HO-I predictions to perform detector supervision, see Figure 2.

Discrete Variable Sampling in Computer Vision. In this work, we treat HO-I target alignment as a hard-valued, discrete variable sampling: Amongst all possible target-prediction pair(s), which subset(s) should be selected for detector supervision? Such decision is non-differentiable therefore ill-suited in convolutional network training. To that end, we resort to a continuous relaxation procedure named Gumbel-Softmax trick, which allows end-to-end training via discrete variables [17, 25]. Gumbel-Softmax has successfully been used to sample convolutional layers [30], filters [7] or channels [3].

In this work, we adapt Gumbel-Softmax to select the target HO-I for detector supervision.

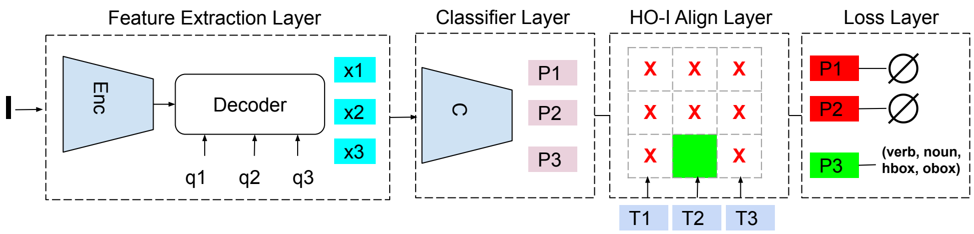

Figure 3: Align-Former consists of four main layers. **Feature Extraction Layer** is an Encoder-Decoder-based visual-transformer that extracts a set of human-object features x_i using the positional queries q_i. Then, **Classifier Layer** generates HO-I predictions P in the form of human-object bounding boxes and verb-noun classes. **HO-I Align Layer** compares HO-I predictions P with potential HO-I targets T to find few-matching pair(s) that are used for HO-I detector supervision using **Loss Layer**.

3 Align-Former for HO-I Detection

Method Overview. An overview of our technique is presented in Figure 2-3. The goal of our network $g_\theta(\cdot)$ is to produce HO-I prediction tuples given an image I as $I \xrightarrow{g_\theta(\cdot)} t'$. Here, HO-I prediction is of size P and represented via $t' = (h', o', v', n')$, where $(h' \in \mathbb{R}^{P\times 4}, o' \in \mathbb{R}^{P\times 4})$ are human-object bounding box predictions, and $(v' \in \mathbb{R}^{P\times V}, n' \in \mathbb{R}^{P\times N})$ are verb-noun class predictions for V verbs and N nouns.

Then, assume we have access to a set of HO-I targets of size T with the same structure $t = (h \in \mathbb{R}^{T\times 4}, o \in \mathbb{R}^{T\times 4}, v \in \mathbb{R}^{T\times V}, n \in \mathbb{R}^{T\times N})$. To supervise Align-Former, we propose to minimize the following objective:

$$\min_\theta (A \times t, t') \tag{1}$$

where we omit θ from now on for clarity. A is a binary matrix of size $P \times T$ where only few entries are non-zero. A is applied separately on all tuple members, as $A \times t = (A \times h, A \times o, A \times v, A \times n)$. Here, $A(i, j) = 1$ means prediction i matches (*i.e.*, aligns) with target j to use in supervision. Similarly, $A(i, j) = 0$ indicates target i should not be used in detector supervision. To identify which target-prediction pairs should be used in detector supervision, we rely on geometric and visual priors detailed later.

Finally, replacing t' with $g(I) = C(Dec(Enc(CNN(I)), Q))$ yields:

$$\min(A \times t, C(Dec(Enc(CNN(I)), Q))) \tag{2}$$

which is detailed in four Sections:

- **HO-I Align Layer (§3.1)** generates the alignment matrix A that pairs few HO-I prediction(s) with HO-I target(s),
- **Classification Layer (§3.2)** generates human-object bounding boxes and verb-noun classification via $C(x)$ using human-object features x,
- **Feature Extraction Layer (§3.3)** generates features via $x = Dec(Enc(CNN(I)), Q)$ via positional queries Q using Encoder-Decoder architecture,
- **HO-I Loss Layer (§3.4)** computes the human-object box and verb-noun classification losses to supervise the detector with the generated HO-I targets t.

3.1 HO-I Align Layer

HO-I align layer consists of two sub-layers, *i)* Prior layer that judges the compatibility between all HO-I targets and predictions, *ii)* Discretization layer that binarizes the likelihood values to obtain the final hard-alignment.

3.1.1 Discretization Layer

Assume we are given a scoring function $S \in \mathbb{R}^{P \times T}$ where $S(i,j)$ encodes how compatible HO-I prediction t_i' and HO-I target t_j matches. Our goal is to discretize this matrix to obtain the final hard-valued alignment decision.

To perform this, we discretize S such that only few members will be non-zeros. Specifically, given raw values of S, we apply the following operation:

$$A = \sigma(S+G) \geq \delta \tag{3}$$

where $\delta = 0.5$ is the hard-threshold value, G is the Gumbel noise [17, 25] added to the matrix S for regularization, and $\sigma(\cdot)$ is the sigmoid activation to bound S between $[0,1]$. Note that Gumbel-noise is crucial to avoid any degenerate solutions like all 1s.

This operation yields the binary alignment matrix $A \in \{0,1\}$ where only a few entries are non-zero.

3.1.2 Prior Layer

To compute the compatibility between HO-I targets & predictions, we resort to a convex combination of geometric and visual priors as $S = \alpha_g * GP + \alpha_v * VP$. Our intuition is that for an HO-I target to be a good candidate for detector supervision, it needs to be compatible both in terms of human-object bounding boxes (geometric) and verb-noun classes (visual).

Geometric Prior $GP(\cdot)$ computes the bounding box compatibility of human-objects via L_1 distance as:

$$GP = \exp\left(-\frac{\sum_{ij} ||h_i' - h_j|| + ||o_i' - o_j||}{\tau}\right) \tag{4}$$

where the exponential function $\exp(\cdot)$ converts the distance values to similarity where $\tau = 1$.

Visual Prior $VP(\cdot)$ computes how well a given target-prediction pair matches in terms of HO-I classes. Remember that our HO-I targets enumerate existing HO-I from the image in terms of verb-noun pairs. Therefore, $VP(\cdot)$ is calculated as:

$$VP = v' * v^T + n' * n^T \tag{5}$$

where verb-predictions are of size $v' \in \mathbb{R}^{P \times V}$ and verb-targets are of size $v \in \mathbb{R}^{T \times V}$ for V distinct verbs. Similarly, noun-predictions are of size $n' \in \mathbb{R}^{P \times N}$ and noun-targets $n' \in \mathbb{R}^{T \times N}$ for N distinct nouns.

3.2 HO-I Classification Layer

Classifier layer is responsible for generating HO-I predictions t' consisting of human-object bounding box predictions (h', o') as well as verb-noun category predictions (v', n').

Human-Object Bounding Box Classifiers are two multi-layer perceptrons $g^h(\cdot)$ and $g^o(\cdot)$ that maps human-object features x to coordinates as $(h', o') = (\sigma(g^h(x)), \sigma(g^o(x)))$.

Verb-Noun Classifiers are also two multi-layer perceptrons as $g^v(\cdot)$ and $g^n(\cdot)$ that learns to map human-object features x to corresponding verb-nouns as $(v', n') = (\sigma(g^v(x)), (g^n(x)))$.

3.3 HO-I Feature Extraction Layer

Our backbone needs to encode: *i)* Object-object relations, *ii)* Relative object positions that are critical to perform HO-I alignment and detection. To that end, we implement the feature extractor as a visual-transformer based on DETR [4]. The feature extractor yields human-object features $x \in \mathbb{R}^{P \times D}$, and consists of three sub-layers: Backbone, Encoder and Decoder, which are detailed below.

Backbone ($x = CNN(I)$). Backbone is a deep CNN [15] that extracts global feature maps from the input image I of size $x \in \mathbb{R}^{H \times W \times C}$ where $[H, W]$ are the height-width of the feature map, and C is the number of channels.

Encoder ($x = Enc(x)$). Encoder further processes the global feature map from the backbone to increase positional and contextual information. We first reduce the number of channels from the backbone to a much smaller size via 1×1 convolutions of $C \times D$. Then, the resulting feature map $\mathbb{R}^{H \times W \times D}$ is collapsed in the spatial dimension as $\mathbb{R}^{D \times HW}$ where each pixel becomes a "token" represented by D dimensional features. Finally, this feature undergoes a few self-attention operations via few multi-layer perceptrons, residual operations, and dropout. At each step, pixel positions are added to the feature map to retain position information.

Decoder ($x = Dec(x, Q)$). The Decoder is a combination of self-attention and cross-attention layers, which yields the final human-object features. The Decoder takes as input the Encoder output $x \in \mathbb{R}^{D \times HW}$ as well as fixed positional query embeddings $Q \in \mathbb{R}^{P \times D}$. Decoder alternates between the cross-attention between the feature map x and Q, as well as self-attention across queries. Cross-attention extracts features from the global feature maps, whereas self-attention represents object-object relations necessary for HO-I detection. Decoder is implemented as multi-layer perceptrons. Final output is $x \in \mathbb{R}^{P \times D}$ that encodes positional and appearance-based representations of potential human-object pairs within the image.

3.4 HO-I Loss Layer

Our loss function ensures that the predicted human-object bounding boxes as well as the verb-noun predictions are in line with the aligned HO-I targets.

The loss function $\mathcal{L}$ is a composite of bounding box, classification, and sparsity losses as $\mathcal{L} = \mathcal{L}_{box} + \mathcal{L}_{class} + \mathcal{L}_{sparse}$. Here, $\mathcal{L}_{box}$ computes the L_1 distances between human-object predictions and (aligned) targets as $\mathcal{L}_{box} = \mathcal{L}_{human} + \mathcal{L}_{object}$. And, $\mathcal{L}_{class} = \mathcal{L}_{verb} + \mathcal{L}_{noun}$ are implemented via classical cross-entropy. As there can be multiple verbs for each instance, we use sigmoid activation before computing the verb loss.

Sparsity Loss. Finally, sparsity loss minimizes $\mathcal{L}_{sparse} = \frac{1}{P \times T} \sum_{ij} A_{ij}$ where $\frac{1}{P \times T}$ is a constant normalizing factor to bound the loss. This ensures the sum over all entries within the alignment matrix A is minimized, leading to only few pairs of HO-I predictions and targets to be aligned for further supervision.

Implementation. We set the number of predictions as $|P| = 100$. Our network is implemented using PyTorch [26]. Feature size D from the last layer of the Decoder is set to $D = 256$. Both human-object bounding box classifiers and verb and noun predictors are 2-layer perceptrons with ReLU activation in between.Initial learning rate is set to 10^{-6} for the ResNet backbone and 10^{-5} for the rest of the parameters. We use weight-decay to regularize the network with 10^{-4}. We train the network for 150 epochs with an effective batch size of 16 over 8 GPU Titan cards. We decay the learning rate linearly with $10^{-}1$ after epoch 100.

4 Experimental Setup

Datasets. We experiment on two large-scale standard datasets, namely HICO-DET [5] and V-COCO [13]. *i) HICO-DET* contains 38k images for training and 9.6k images for testing. Images contain 117 distinct verbs and 80 distinct nouns together, making 600 `<verb, noun>` pairs. For each noun, there exists a case of "no-interaction", where at least a single human and the target object is visible, even though not interacting. We only use HO-I alignment annotations for testing, and not training, since our goal is to evaluate HO-I detection via image-level supervision. *ii) V-COCO* builds upon MS-COCO [23] where the authors annotate subset of images with human-object alignments and their (inter-)action. The type of interactions is riding, reading and smiling. The dataset exhibits 2.5k images for training, 2.8k images for validation, and 4.9k images for testing.

Metric. We use the mean Average Precision (mAP) metric for evaluation as is the standard [5, 13]. A human-object interaction is true positive only if both humans and objects have an Intersection-over-Union with a ground-truth HO-I pair above > 0.50 *and* they are assigned to the correct interaction categories.

Evaluation. *i) HICO-DET:* We use the evaluation code presented in the server [2]. We compute the mean over all three splits of full, rare, and non-rare in HICO-DET. We provide comparison on three standard splits. *Full*: All 600 categories, *Rare*: 138 categories with less than or equal to 10 training instances, *Non-Rare*: 462 categories with more than 10 training instances. *ii) V-COCO:* We use the evaluation code presented in authors' code [1]. We evaluate using three different standard scenarios. *Agent*: We report the human interactor detection performance, *Scenario-1*: We report the detection of humans and objects together, *Scenario-2*: We report the detection of humans and objects where the object predictions for object-less interactions (*i.e.*, smiling) is ignored.

Baselines. We compare Align-Former to *i) Weakly-supervised HO-I detectors*: PPR-FCN [31] and MX-HOI [21] that performs HO-I detection without alignment supervision. *ii) Strongly-supervised variants*: To measure the upper bound performance as a reference, we also report MX-HOI and Align-Former performance via strong alignment supervision.

5 HO-I Detection on HICO-DET & V-COCO

5.1 Comparison to The State-of-The-Art

Method	Backbone	Alignment-Supervised?	Full	Rare	Non-Rare
PPR-FCN [31]	ResNet-101	✗	15.14	10.65	16.48
MX-HOI [21]	ResNet-101	✗	16.14	12.06	17.50
Align-Former (ours)	ResNet-50	✗	<u>19.26</u>	<u>14.00</u>	<u>20.83</u>
Align-Former (ours)	ResNet-101	✗	**20.85**	**18.23**	**21.64**
MX-HOI [21]	ResNet-101	✓	17.82	12.91	19.17
Align-Former (ours)	ResNet-50	✓	25.10	17.34	27.42
Align-Former (ours)	ResNet-101	✓	27.22	20.15	29.57

Table 1: Human-Object Interaction Detection mAP on HICO-DET [5]. Our method outperforms existing techniques over all splits of full, rare, and non-rare.

HICO-DET Results are presented at Table 1. Overall, Align-Former outperforms the other two techniques by 3.12 mAP via ResNet-50 and 4.71 mAP via ResNet-101 on all categories. This confirms that HO-I detection benefits from the end-to-end alignment of the targets and the predictions. Our improvement is even more pronounced on the rare split via 6.17 mAP using ResNet-101, exhibiting the sample efficiency of our technique.

Method	Backbone	HICO-DET Pre-Trained?	Alignment-Supervised?	Agent	Scenario 1	Scenario 2
Align-Former	ResNet-50	✗	✗	24.63	13.90	14.15
Align-Former	ResNet-50	✓	✗	<u>27.95</u>	<u>15.52</u>	<u>16.06</u>
Align-Former	ResNet-101	✗	✗	20.00	10.44	10.79
Align-Former	ResNet-101	✓	✗	**30.02**	**15.82**	**16.34**
Align-Former	ResNet-50	✗	✓	66.78	50.20	56.42
Align-Former	ResNet-101	✗	✓	68.00	55.40	62.15

Table 2: Human-Object Interaction Detection mAP on V-COCO [13]. Even though the performance is limited when trained from scratch on V-COCO, HICO-DET pre-training yields a considerable improvement on V-COCO.

V-COCO Results are presented at Table 2. We only compare to our own baselines [2]. We evaluate two different settings. *i) Training on V-COCO from scratch*: Since the number of training images are quite limited (only 2k examples), training on V-COCO without alignment supervision yields limited accuracy on all three settings. *ii) Transfer learning from HICO-Det*: where we fine-tune a HICO-DET pre-trained model on V-COCO. In all cases, pre-training on HICO-DET helps significantly. As one of the major goal of annotation-free training is the ability to pre-train on large-scale benchmarks, we see this as a promising direction in HO-I detection with cheap image-level supervision.

We confirm that our model yields competitive performance on HICO-DET against competing benchmarks on all full, rare and non-rare splits, and showcases promising first results without alignment supervision on V-COCO, especially via transfer learning.

[2]Neither of the existing baselines (PPR-FCN and MX-HOI) evaluates on V-COCO. Additionally, strongly supervised stream of MX-HOI (No-Frills HO-I [14].) also is not evaluated on V-COCO

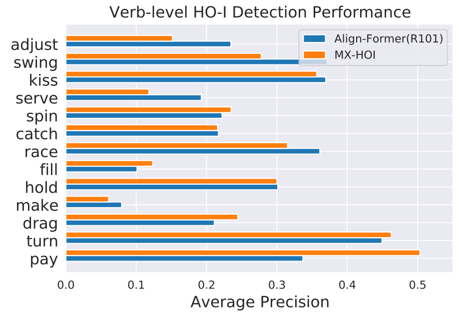

Figure 4: Verb-level Performance on HICO-DET [5]

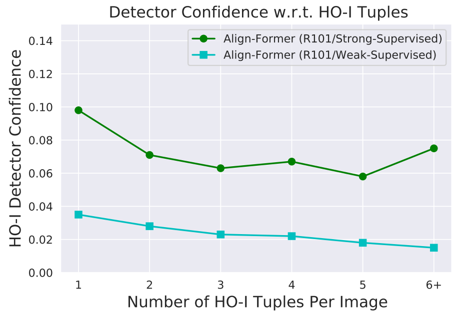

Figure 5: HO-I detector confidence w.r.t. number of HO-I tuples in an image on HICO-DET [5].

5.2 Further Analysis

In this section, we provide analysis to better understand the contribution of Align-Former.

Verb-level Performance Comparison. We visualize verb-level performance difference between weakly supervised Align-Former and MX-HOI in Figure 4. We observe that Align-Former outperforms for pose and part-driven interactions like adjust, swing or kiss, while underperforming for scene-driven interactions like pay or turn. This indicates end-to-end learning of pose-based representations is more valuable than hand-crafted pose representations as in MX-HOI. For more results, refer to our Supp. material.

W/ *vs.* W/O Alignment Supervision. To better understand the gap between strongly *vs.* weakly supervised HO-I detection, we provide results of MX-HOI with strong supervision on HICO-DET in Table 1 as well as strongly supervised Align-Former in both datasets (Table 1- 2). Our method is flexible as it can be easily trained with strong and weak supervision with no change in architecture, whereas MX-HOI ensembles two CNNs (a weak [31] and strong [14] CNN) to do so.

We have three main findings. *i)* Weakly-supervised Align-Former outperforms strongly supervised MX-HOI on HICO-DET (Table 1), which indicates our method compensates for the lack of supervision with its representational power. *ii)* Strongly supervised Align-Former outperforms weakly supervised Align-Former on both datasets (Table 1- 2). This shows Align-Former better leverages the supervision when is used, and there is a room for improvement in weakly-supervised techniques. *iii)* In Figure 5, we plot the confidence of strongly *vs.* weakly supervised Align-Former as a function of number of HO-I tuples in an image on HICO-DET. As can be seen, strongly-supervised variant retains its performance whereas weakly-supervised degrades in confidence, which may help explain the performance gap between the two variants of Align-Former.

ResNet-101 vs. ResNet-50. We implement Align-Former with ResNet-50 and 101. Even though we do not observe significant difference at the verb- or object- level, the difference is at the interaction-level. Our findings are: *i)* ResNet-101 outperforms $ResNet-50$ on both datasets across all settings, *ii)* Surprisingly, ResNet-101 outperforms especially on the rare split of HICO-DET, and exhibits better transferability to V-COCO, despite higher number of parameters.

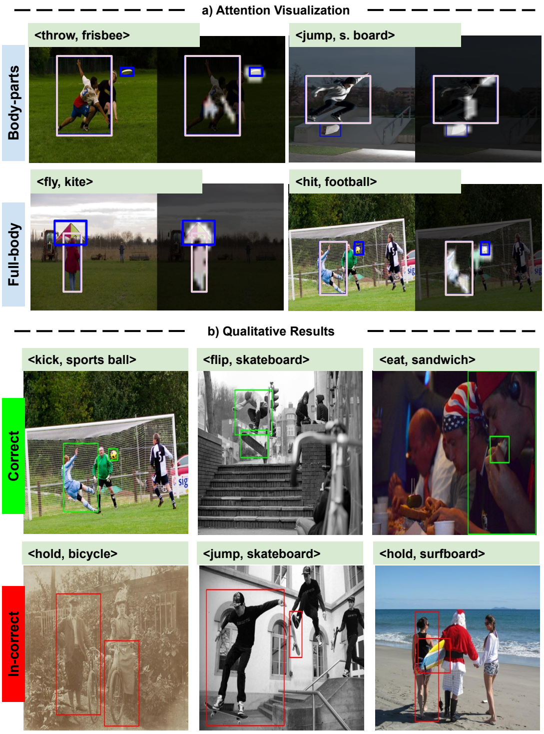

Figure 6: *a)* Attention analysis of Align-Former reveals the focus on body-part and full-body. *b)* Qualitative analysis of Align-Former reveals it can detect both dynamic and static interactions.

Qualitative Inspection. *i) Attention Analysis*: To understand where Align-Former is looking at to perform HO-I alignment and detection, we present the attention matrix for a set of queries from the last layer of the Decoder in Figure 6-(a). We observe that Align-Former attends on body-parts when the visual information is sufficient, and full-body when the human-object has small scale. *ii) Qualitative Results*: Finally, we visualize high-confident detection examples in Figure 6-(b). We observe that Align-Former can detect both dynamic interactions like `<kick, sports ball>` or static interactions like `<eat, sandwich>`. However, our method fails when humans can not be paired with their object of interaction, as is visualized in the bottom row.

6 Conclusion

This paper addressed HO-I detection from images. We proposed Align-Former, a visual-transformer based CNN that can learn to detect HO-I without alignment supervision, via image-level supervision. We equip Align-Former with HO-I align, a novel layer that learns to select correct detection targets based on geometric and visual priors. We show that Align-Former outperforms existing techniques for HO-I detection on HICO-DET especially on rare HO-I, and yields promising results on V-COCO, confirming the efficacy of our method. We hope our work inspires future research on reducing supervision in HO-I detection.

Acknowledgements. We are grateful for the constructive and useful feedback from all our anonymous reviewers as well as the meta reviewer, which helped us to greatly improve our manuscript.

References

[1] Vcoco evaluation server. In *Link*, 2015.

[2] Hico-det evaluation server. In *Link*, 2018.

[3] Babak Ehteshami Bejnordi, Tijmen Blankevoort, and Max Welling. Batch-shaping for learning conditional channel gated networks. *arXiv preprint*, 2019.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

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

[6] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *CVPR*, 2021.

[7] Zhourong Chen, Yang Li, Samy Bengio, and Si Si. You look twice: Gaternet for dynamic filter selection in cnns. In *CVPR*, 2019.

[8] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. *NeurIPS*, 2016.

[9] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *BMVC*, 2018.

[10] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *ECCV*, 2020.

[11] Ross Girshick. Fast r-cnn. In *ICCV*, 2015.

[12] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018.

[13] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint*, 2015.

[14] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, appearance and layout encodings, and training techniques. *arXiv preprint*, 2018.

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

[16] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, 2020.

[17] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint*, 2016.

[18] Mert Kilickaya and Arnold WM Smeulders. Structured visual search via composition-aware learning. In *WACV*, 2021.

[19] Mert Kilickaya, Noureldien Hussein, Efstratios Gavves, and Arnold Smeulders. Self-selective context for interaction recognition. *arXiv preprint arXiv:2010.08750*, 2020.

[20] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. In *ECCV*, 2020.

[21] Suresh Kirthi Kumaraswamy, Miaojing Shi, and Ewa Kijak. Detecting human-object interaction with mixed supervision. In *WACV*, 2021.

[22] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *CVPR*, 2020.

[23] 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.

[24] Yang Liu, Qingchao Chen, and Andrew Zisserman. Amplifying key cues for human-object-interaction detection. In *ECCV*, 2020.

[25] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint*, 2016.

[26] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[28] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[30] Andreas Veit and Serge Belongie. Convolutional networks with adaptive inference graphs. In *ECCV*, 2018.

[31] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn. In *ICCV*, 2017.