# Human-Object Interaction Detection *without* Alignment Supervision

## Supplementary Material

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

Visual Sensing Lab
University of Amsterdam
Amsterdam, Netherlands

## 1 Implementation Details

**Prior Layer Coefficients.** Prior layer computes a weighted sum of visual prior and geometric prior as $S = \alpha_g * GP + \alpha_v * VP$, where we simply set them equal as $\alpha_g = 0.5, \alpha_v = 0.5$.

**Augmentation.** During training, we resize the smallest side of the input image to different scales as $\{480, 624, 784, 800\}$. During testing, we use the single scale of 800. We apply random cropping where the smallest image dimension is 600 prior to resize. We remove the HO-I targets that fall outside the cropped image (if any). Lastly, we use random horizontal flips.

**HO-I Target Generation** $(t)$**.** Remember that our goal is to detect HO-I without alignment supervision between humans-objects and interactions. To that end, we first sample human and object regions from the respective dataset [1, 2], as well as the interaction list. Then, we exhaustively pair all humans and objects and humans with other humans. To create interaction annotations, we simply repeat the list of existing HO-I at the image-level for all candidate targets. One can refer to Figure-2 within the main paper for three target HO-I examples.

## 2 Further Analysis

**Verb-level Performance Comparison.** First, we visualize verb-level performance comparison to MX-HOI [3] in Figure 1. Our observations are the following:

- The improvement of Align-Former is generic, as it improves over most of the verbs.
- The improvement is even more pronounced for pose-driven interactions like blowing, hugging, kissing, or adjusting, indicating that end-to-end learning of the pose is a better alternative to hand-crafted pose stream in MX-HOI [3].
- All three techniques perform poorly for the case of no-interaction, indicating no-interaction is hard to learn without strong HO-I alignment supervision.

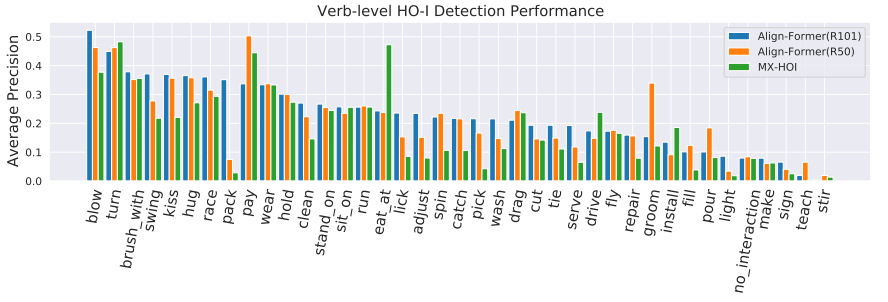


Figure 1: Verb-level average precision for HO-I detection on HICO-DET [1]. We observe that the contribution of Align-Former is generic across different verbs.

**Performance Distribution Over Nouns.** We present the performance comparison across different nouns on HICO-DET [1] in Figure 2. As can be seen, the improvement of Align-Former is generic across different noun groups. In addition, the biggest improvement is obtained on sport objects like skateboard, snowboard, indicating that Align-Former can induce pose-related visual features necessary to recognize sport interactions.

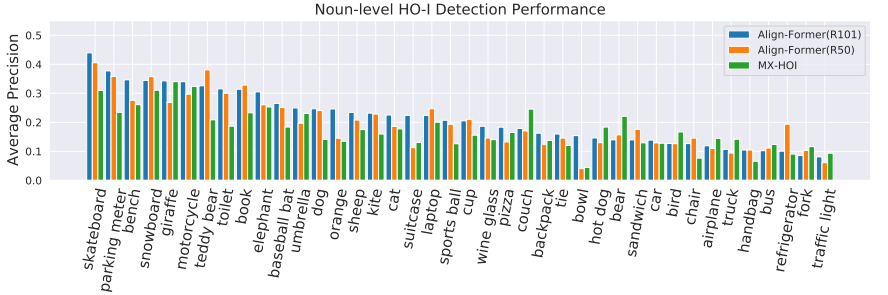


Figure 2: Noun-level average precision for HO-I detection on HICO-DET [1]. We observe that the contribution of Align-Former is generic across different nouns.

**Attention Analysis.** We visualize additional attention results from the last decoder layer of Align-Former, for some queries in Figure 3. Our method attends both on local body-parts and global full body. *Body-parts*: In local body-parts, active human body regions such as hands, upper arms and upper legs are activated, as well as interacted object regions. *Full-body*: When the human-object has low visual scale, such as flying kite, the network aggregates information from all over the human and object regions.

**Qualitative Results.** We visualize additional qualitative detection results in Figure 4. *Correct*: Our network can successfully detect HO-I, even when there are multiple interactors, such as the group of bicycle riders. *In-Correct*: Our network fails to detect when the target noun is mis-labelled (*i.e.*, bottle vs. hair drier) or the target verb is mis-labelled (*i.e.*, sitting vs. lying on). *Un-annotated*: An interesting case is where our network makes a correct prediction, such as carrying a handbag or not interacting with the bicycle, however they count

as false detections due to missing annotations. This motivates the need to utilize cheap annotations like image-level HO-I for detection, since annotating each and every HO-I instance from the image is not possible.

**Qualitative Results across multiple HO-I.** Finally, we visualize success and failure cases when there are multiple HO-I within the same image in Figure 5. Our model successfully detects HO-I within the clutter of spectators, however its performance is limited, especially in sport activities motivating future research in this area.

# References

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

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

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


Figure 3: Attention analysis of Align-Former reveals the focus on body-part and full-body.


Figure 4: Qualitative analysis of Align-Former reveals it can detect both dynamic and static interactions.

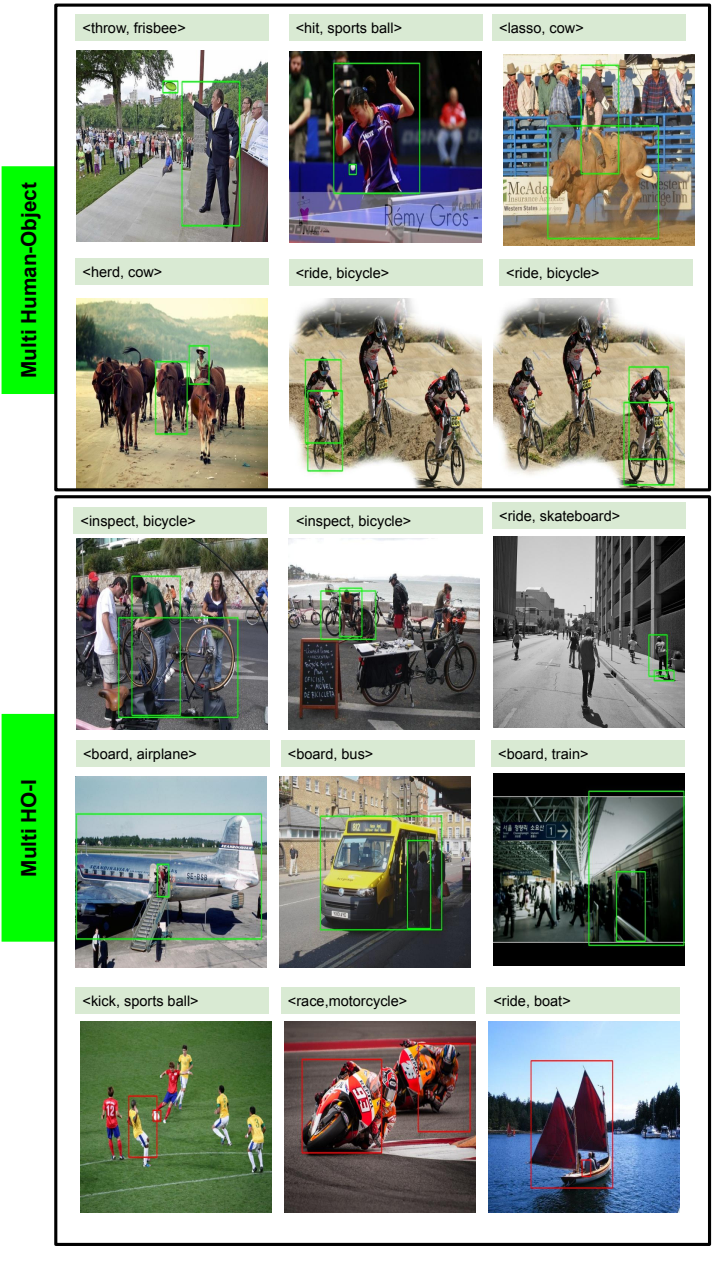


Figure 5: Qualitative analysis of Align-Former when there are multiple HO-I tuples within the image. As can be seen, in some cases, Align-Former can successfully detect HO-I even within clutter. In the first example on top-left, despite there is a big crowd watching and inspecting, the model localizes the interaction of `<throw, frisbee>`. A similar trend follows when multiple people are boarding on vehicles, or riding skateboard. However, in cases like sport activities (*i.e.*, , football or race), our model fails to align human with the corresponding object of interaction.