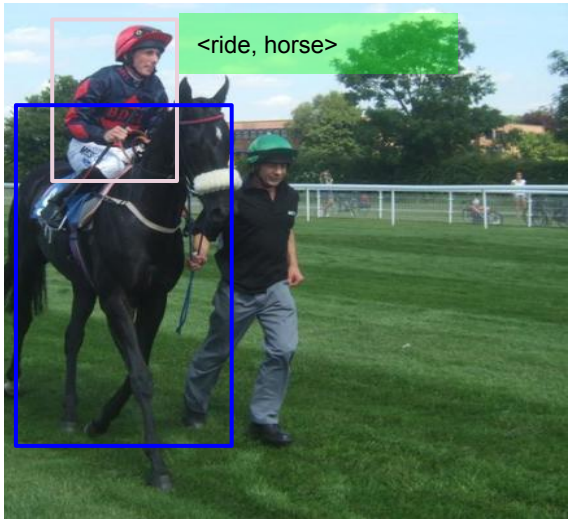
Human-Object Interaction Detection via Weak Supervision

Mert Kilickaya

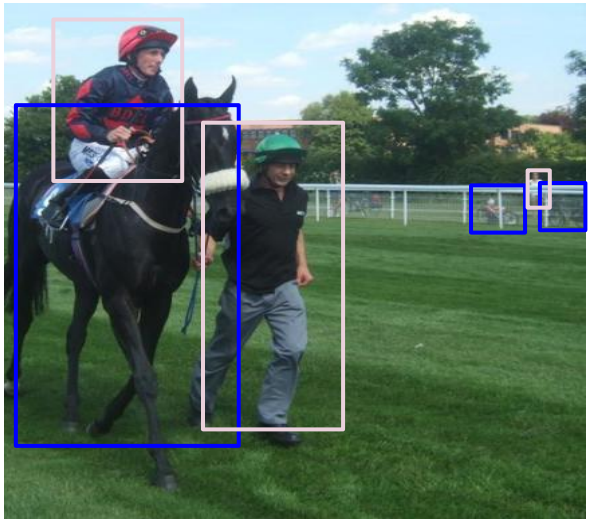
Arnold Smeulders

Human-Object Interaction (HO-I) Detection

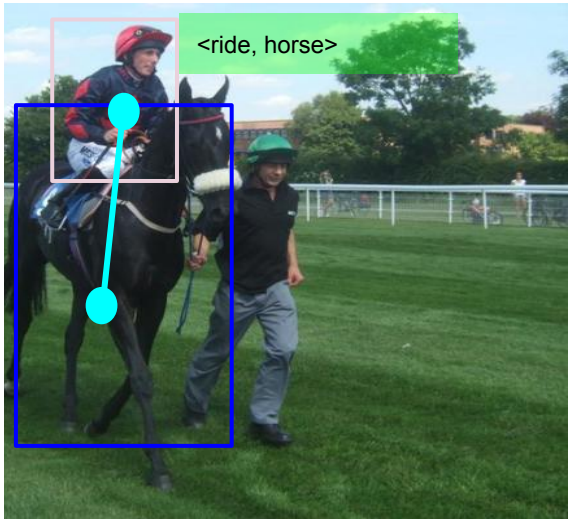


1) Find interacting **<human, object>** boxes, 2) Determine **<verb, noun>** interaction classes.

HO-I Detection via Strong (Instance-level) Supervision



Instance-level Bounding Box

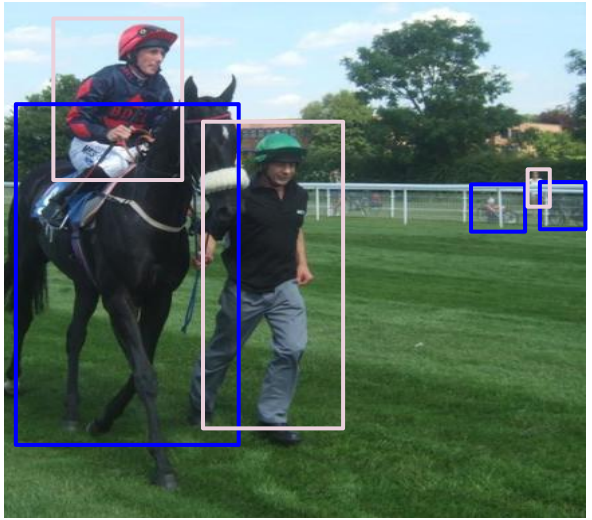


Instance-level Interaction

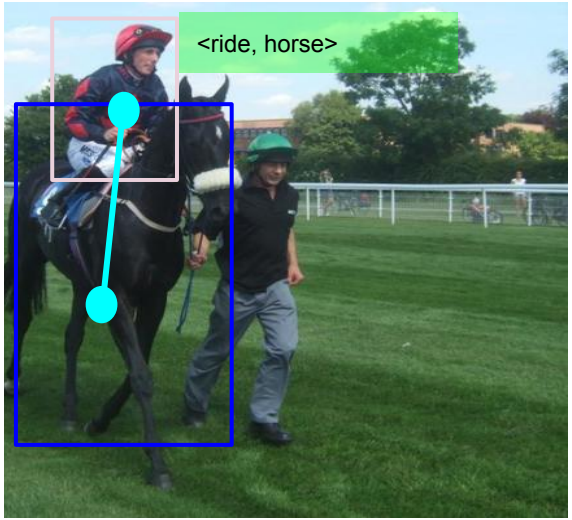
Image-level Interaction

Strong supervision is expensive: **150k instance annotations**.

HO-I Detection via Weak (Image-level) Supervision



Instance-level Bounding Box

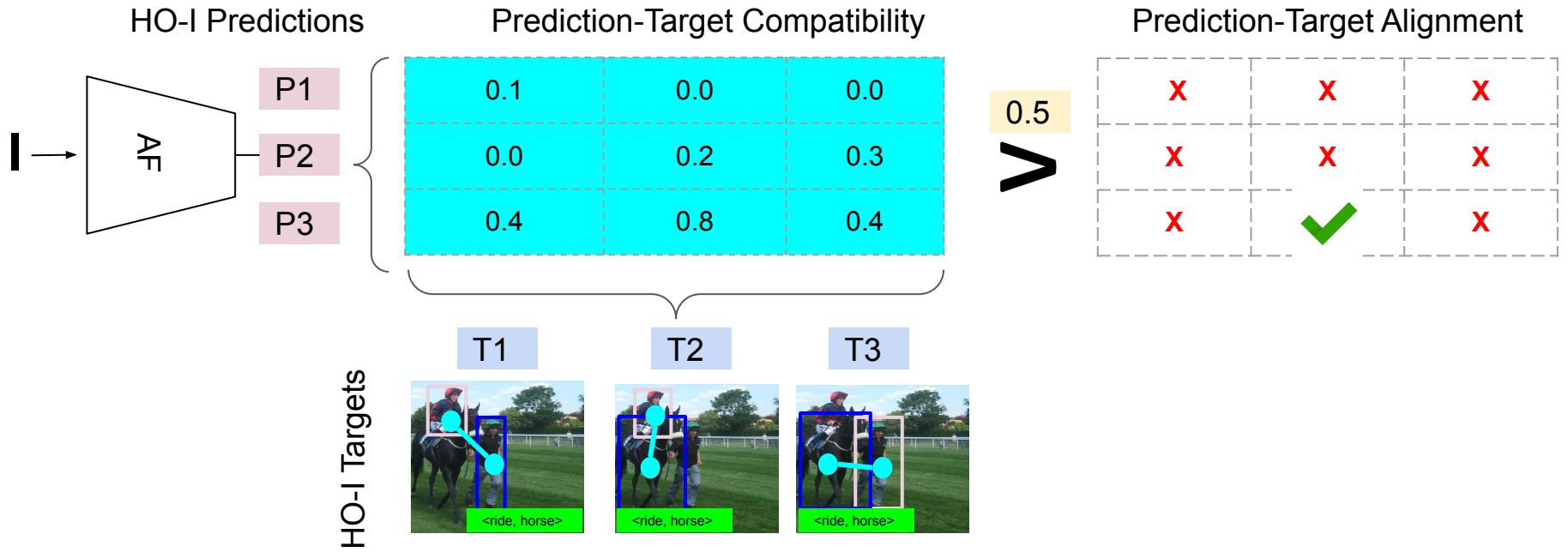


Instance-level Interaction

Image-level Interaction

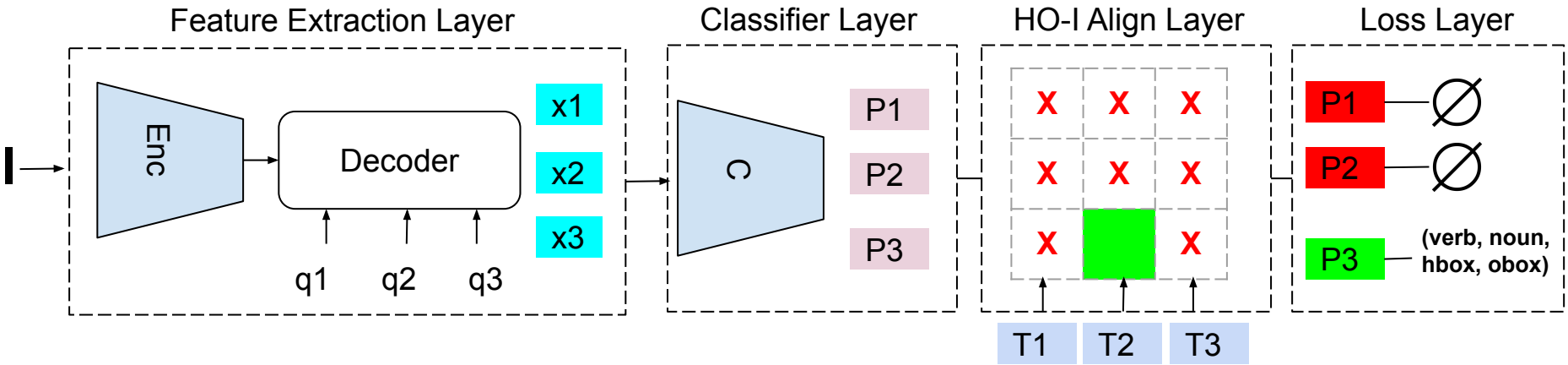
We propose to detect HO-I via much cheaper, weak supervision: Align-Former.

Align-Former: Core Idea



Align-Former: Learn to select <u>few</u> correct detection target(s) to supervise the detector from weak supervision.

Align-Former: Implementation



Align-Former: An end-to-end trained visual Transformer with *non-differentiable* HO-I align operation.

Exp 1: State-of-the-Art Comparison on HICO-DET

HICO-DET: **40k** images | **117** Verb | **80** Objects | **600** HO-I: **480** Non-rare, **120** Rare Interactions

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

Large improvement on all setups, even against supervised variants.

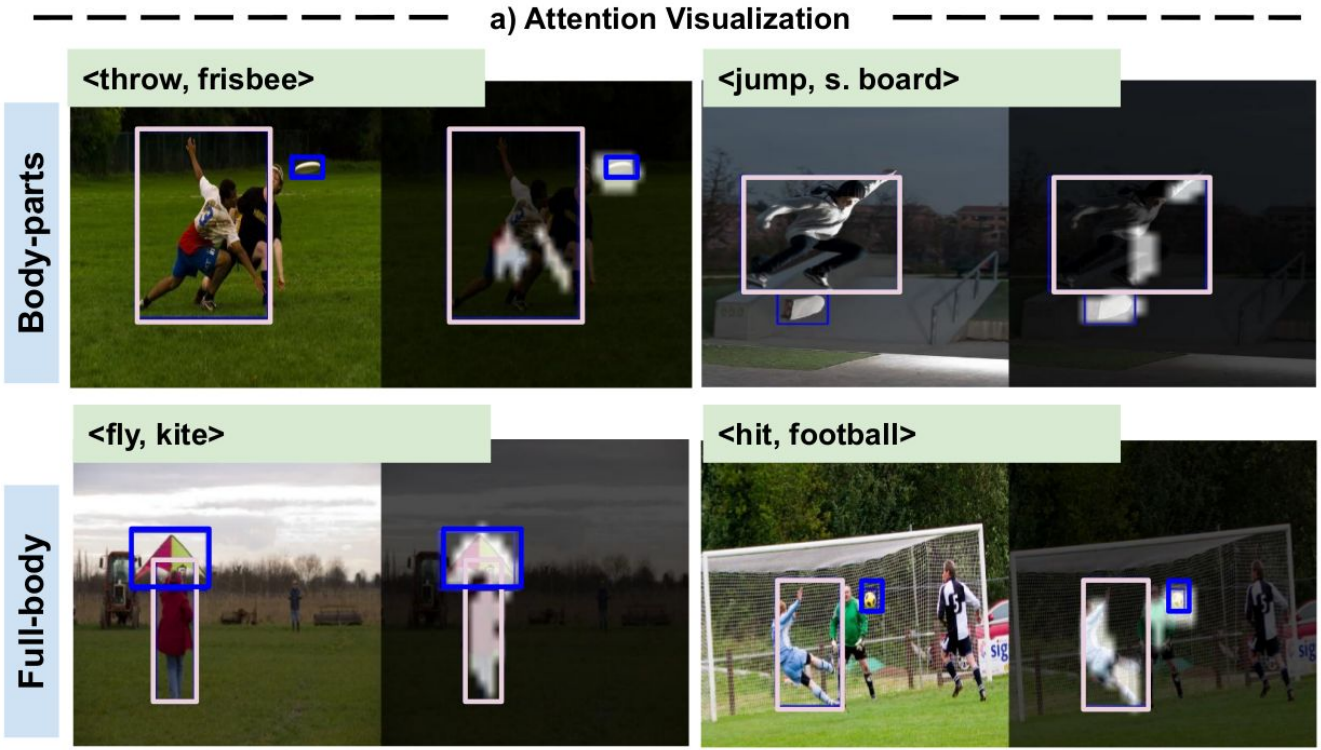
Exp 2: First Results on V-COCO via Weak Supervision

V-COCO: **2.5k** images | **26** Verb | **80** Objects | **120** HO-I

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	✓	✗	27.95	15.52	16.06
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

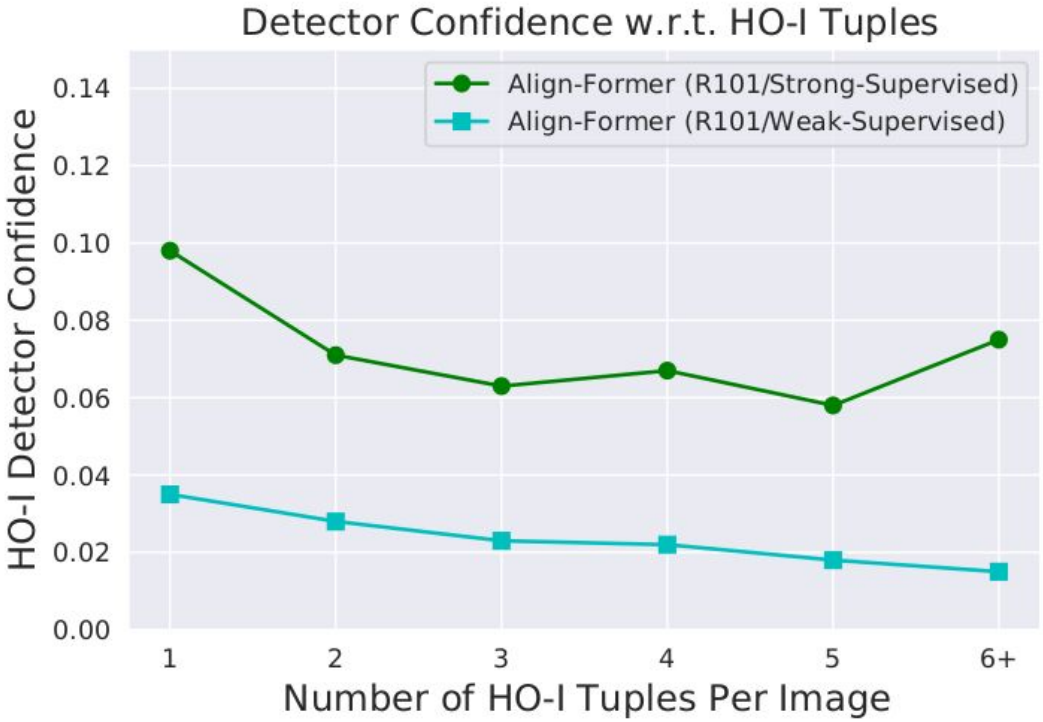
V-COCO: Weakly supervised HO-I detection for the first time with very limited number of images

Align-Former: Attention



Align-Former: Leverages 1) Body-parts, 2) Full-body information for detection.

Align-Former: Limitation



Align-Former: Performance drops drastically with respect to number of human-object pairs.

Summary

HO-I Detection via Weak Supervision: Able to omit 150k HO-I instance annotations.

Align-Former: Able to select correct HO-I detection targets to supervise the underlying detector.

Result: Able to train a sample-annotation efficient detector on: HICO-DET & V-COCO.

Thank you! Any Questions?