Detecting Human-to-Human-or-Object (H^2O) Interactions with DIABOLO

Astrid Orcesi, Romaric Audigier, Fritz Poka Toukam, Bertrand Luvison

Detecting Human-to-Human-or-Object (H^2O) Interactions with DIABOLO

Astrid Orcesi[1,2] Romaric Audigier[1,2] Fritz Poka Toukam[1] Bertrand Luvison[1,2]
[1] Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
[2] Vision Lab, ThereSIS, Thales SIX GTS, Campus Polytechnique, Palaiseau, France
{firstname.lastname}@cea.fr

***Abstract*— Detecting human interactions is crucial for human behavior analysis. Many methods have been proposed to deal with Human-to-Object Interaction (HOI) detection, i.e., detecting in an image which person and object interact together and classifying the type of interaction. However, Human-to-Human Interactions, such as social and violent interactions, are generally not considered in available HOI training datasets. As we think these types of interactions cannot be ignored and decorrelated from HOI when analyzing human behavior, we propose a new interaction dataset to deal with both types of human interactions: Human-to-Human-or-Object (H^2O). In addition, we introduce a novel taxonomy of verbs, intended to be closer to a description of human body attitude in relation to the surrounding targets of interaction, and more independent of the environment. Unlike some existing datasets, we strive to avoid defining synonymous verbs when their use highly depends on the target type or requires a high level of semantic interpretation. As H^2O dataset includes V-COCO images annotated with this new taxonomy, images obviously contain more interactions. This can be an issue for HOI detection methods whose complexity depends on the number of people, targets or interactions. Thus, we propose DIABOLO (Detecting InterActions By Only Looking Once), an efficient subject-centric single-shot method to detect all interactions in one forward pass, with constant inference time independent of image content. In addition, this multi-task network simultaneously detects all people and objects. We show how sharing a network for these tasks does not only save computation resource but also improves performance collaboratively. Finally, DIABOLO is a strong baseline for the new proposed challenge of H^2O-Interaction detection, as it outperforms all state-of-the-art methods when trained and evaluated on HOI dataset V-COCO. We hope that this new dataset and new baseline will foster future research. H^2O is available on https://kalisteo.cea.fr/.**

I. INTRODUCTION

One of the requirements for visual analysis of human behavior is to recognize human actions. Many methods [2], [1], [23], [22] have dealt with human action recognition in video. The goal is to classify a whole video clip containing a single main action. However, substantial video datasets containing multiple simultaneous actions with localization of their subjects and targets are still missing. Other methods [10], [8], [3] have dealt with the so-called Human-Object Interaction (HOI) detection task from a single image. It consists in determining and locating the list of triplets $< subject, verb, target >$ which describe all the simultaneous interactions in an image. Several image datasets for HOI have been made available [12], [4]. However, they focus on targets which are non-human, called "objects" hereafter. Therefore, a lot of human interactions, such as social or violent interactions, are not taken into consideration. To analyze human behavior, Human-to-Human Interactions (HHI), i.e. interactions between people, cannot be ignored and decorrelated from HOI. For example, in video surveillance applications, it is interesting to recognize fighting people, to distinguish them from hugging people, and also to detect people kicking urban equipment. The lack of datasets dealing with both types of human interactions is the first motivation for proposing a new dataset called H^2O (Human-to-Human-or-Object).

Second, verb taxonomies used by existing datasets [12], [4] are sometimes ambiguous. For example, some types of interactions correspond to synonymous verbs (*e.g.*, to inspect or to watch an object, to hold or to carry a cell phone, to read or to look at a book...) which sometimes only differ from the type of target (*e.g.*, to surf, to snowboard, or to skateboard) or requires a high level of semantic interpretation of the context or the intent (*e.g.*, to hold, to pick or to buy an apple, to ride or to sit on a horse). Conversely, some English verbs can merge different types of human postures or attitude that could be distinguished in another language (*e.g.*, to ride a horse or a bus do not correspond exactly to the same body attitude in relation to the target object). This motivates us to introduce a novel taxonomy of verbs, intended to be closer to a description of the human body attitude in relation to the surrounding targets of interaction, and less dependent on the environment, the target type or the linguistic arbitrariness. To constitute H^2O dataset, we re-annotated images from V-COCO dataset [12] with this new taxonomy including both object and human targets, and added new images to enrich the dataset with HHI verbs.

Consequently, this new dataset motivated us to propose a novel method named DIABOLO (Detecting InterActions By Only Looking Once) to address H^2O Interaction (H^2OI) detection problem. Formally, image-based HOI or H^2OI detection can be decomposed in three steps: detecting objects and people (named together "instances" hereafter) in the image, pairing interacting instances, and classifying the interaction. Most state-of-the-art approaches [26], [27], [24], considered as two-stage methods, rely on an external object detector to perform the first step (identify candidates) then feed a second network with the detections to perform pairing and classification. Other methods [14], [7] include all steps in the same multi-task network. However, they dissociate

the learning of the object detection task from the interaction detection task. We propose to study the effect of learning both tasks jointly or consecutively and how these tasks can collaborate to improve overall performance while reducing the computation resource with the use of a shared backbone.

Finally, for many applications such as video surveillance, ambient assisted living or cobotics, the response time of the method is an important criterion in order not to induce a latency in the real-time analysis. Moreover, the image may have a lot of people and objects that possibly interact together. Ideally, computation time of the method should not be affected by the density of the image. However, most of state-of-the-art methods cannot guarantee a constant time, as an interaction estimation network is applied on all possible pairs. These methods are not scalable with the number of visible instances and interactions. Instead, the so-called single-shot methods use a single forward pass of the image in the network, which ensures constant computation time.

The contributions of this paper are the following:

- We propose a new dataset H^2O to train and evaluate models on the novel task of Human-to-Human-or-Object Interaction (H^2OI) detection, not limited to non-human targets.
- We introduce a new taxonomy of verbs intended to be closer to the human body attitude relative to the surrounding targets of interaction, and to be less dependent on the context or environment in which the interactions occur. We strove to avoid synonyms or linguistic bias and to reduce room for high-level interpretation.
- We propose DIABOLO, a single-shot method which detects instances of the image and estimates their interactions in a single forward pass throughout the image. This new subject-centric method runs in fast and constant time independently of the image content, and without needing external object detector.
- DIABOLO is a first yet strong baseline for this new H^2OI challenge. Indeed, when trained and evaluated on a HOI detection challenge like V-COCO, DIABOLO outperforms all existing state-of-the-art methods.
- Finally, ablative experiments on DIABOLO show how multi-task learning can improve performance of interaction detection.

The paper is organized as follows: Related work is introduced in Section II. Sections III and III-D present H^2O dataset and DIABOLO method, respectively. Evaluations of DIABOLO on both HOI and H^2OI detection challenges are presented in Section IV, as well as an ablative study and a comparison with state-of-the-art methods.

II. RELATED WORK

A. HOI datasets

Human behavior understanding has often resulted in action recognition in video clip. A significant number of datasets [2], [1], [23], [22] consist of video clips of few seconds which have to be classified among a set of possible actions. The large dataset of video clips AVA [11] partially introduces spatial and temporal localization of the actions: only person bounding boxes with related actions are annotated in the central frame of each clip. No information is provided about the targets of the actions.

More recently, some image datasets focus on relationship between objects in an image. For instance, Visual Genome [15] proposes 108,000 images annotated with 18 visual relationships. The relationships can involve any two objects of the image. But predicates have rather low semantic levels and mostly concern positional relationships (as "behind" or "next to"). This dataset is usually used to study visual relationship, but not human interactions. HCVRD [30] is a subset of Visual Genome which selects human-centered relationships expanded into 927 categories, including actions, (pre)positional and comparative relations. This large number of interactions involves that most of them appear less than 10 times. HICO [5] is a dataset for HOI classification: the images are subject-centric and the whole image corresponds to a list of non-localized interactions. HICO-DET [4] uses 47,776 images from HICO and adds bounding boxes annotations to create an HOI detection dataset. HICO-DET contains 117 verbs some of which are very specific to a target object category or synonyms. Finally, V-COCO [12] is a subset of 10,346 images from COCO [19] annotated with 26 interaction verbs, but the interaction target is limited to a single object per verb.

Regarding HHI datasets, TVHI [13] proposes only few social interactions. [6] focuses only on relative motion between people. None of these datasets propose to merge HOI and HHI. Therefore, we define a new taxonomy of 51 verbs (cf. Section III-B) including both HOI and HHI and avoiding synonyms and target specificities, to be less dependent on the environment and closer to the human body attitude relative to the surrounding targets of interaction. Then we propose the H^2O dataset that includes V-COCO images re-annotated with this taxonomy and 3,666 new images collected to enrich HHI.

B. HOI detection methods

HOI detection in images consists in localizing interacting instances in the scene and pairing them with a specific interaction verb to create a list of triplets $< subject, verb, target >$ of all the visible interactions.

On the one hand, most of the previous methods adopt a two-stage strategy [10], [8], [3], [16], [29], [26], [27], [20], [24]. During the first stage, they use an external object detector as [9] to point out interacting candidates, and then, during the second stage, another network is dedicated to estimate interactions between the proposals. First works are essentially based on the appearance of the objects: [10], [8], [26] extract features from the object location to classify the interactions. More recently, improvement in the second stage have been proposed by using additional information in the image. For instance, [16], [27], [20] use human pose to have a finer analysis of the posture of the interaction subject. Other methods add word embedding [24], [21] or segmentation [29]. Finally, [24] combines all these kinds of additional

information. The major drawback of these methods is that they analyze each possible $< subject, object >$ pair in order to determine all the interactions in the image. Their computation time is therefore quadratic with the number of instances in the image. Some methods provide an alternative to studying all possible pairs and thus accelerate the inference time. [17], [28] model the interaction detection problem as a keypoint detection one. [3] densely estimate interactions over an anchor grid to have computation time independent of the number of instances in the image, then use an external object detector to point out the anchors that actually correspond to instances.

On the other hand, some recent works propose one-stage HOI detectors. [14], [7] only rely on regression and classification to predict interactions. They both integrate an instance detection branch similar to classical object detectors. Interaction detection is then based on the union of regressed bounding boxes [14]. [7] notes that it is better to focus at regions of interaction rather than the whole union box which presents too much unnecessary information. Therefore, [7] proposes an interaction region-centric branch to detect interaction. These two methods initialize their internal object detector with a pre-trained model and then freeze these weights to learn the interaction branch.

Our method is one-stage and contrary to [7], DIABOLO is subject-centric and uses embedding to pair interacting instances. Unlike [24], [21], [29], we do not use any additional information. Finally, DIABOLO is the first multi-task network which trains object and interaction detections jointly.

III. PROPOSED DATASET

In this section, we present H^2O Dataset, an image dataset annotated for Human-to-Human-or-Object interaction detection. We first present the modalities to constitute the dataset, the taxonomy chosen to annotate interactions and compare it to currently available datasets. Then, we present metrics to evaluate algorithm performance.

A. H^2O composition

H^2O is composed of the 10,301 images from V-COCO [12] images to which are added 3,666 images selected in the wild as for COCO dataset [19] and which mostly contain interactions between people. Thus, unlike current available datasets, H^2O presents interactions between human and object but also human and human. As for object annotations, all interacting instances are annotated with bounding boxes, even if they do not belong to the 80 classes of COCO [19]. In total, instances are distributed in 214 classes. However, out of a total of 128,969 annotated instances (58,225 people and 70,744 objects), 96% are part of the 80 classes of COCO. Interactions are exhaustively annotated for each person, whether they are with an object or another person. These annotations were made with Pixano[1] annotation tool.

[1]https://pixano.cea.fr

B. H^2O taxonomy

To annotate H^2O, we defined a new taxonomy of verbs including both HOI and HHI. This taxonomy (cf. Figure 1) is intended to be closer to the human body attitude relative to the surrounding targets of interaction, and less dependent on the environment in which the interactions occur, the target type or the linguistic bias. So we strive to avoid synonymous verbs when their use highly depends on the target type or verbs which require a high level of semantic interpretation.

H^2O dataset is annotated with 51 verbs divided into five categories: (i) verbs describing the general posture of the subject, (ii) verbs related to the way the subject is moving, (iii) verbs used for interactions with objects, (iv) verbs describing human-to-human interactions and finally (v) verbs of interactions involving strength or violence which can affect either objects or people.

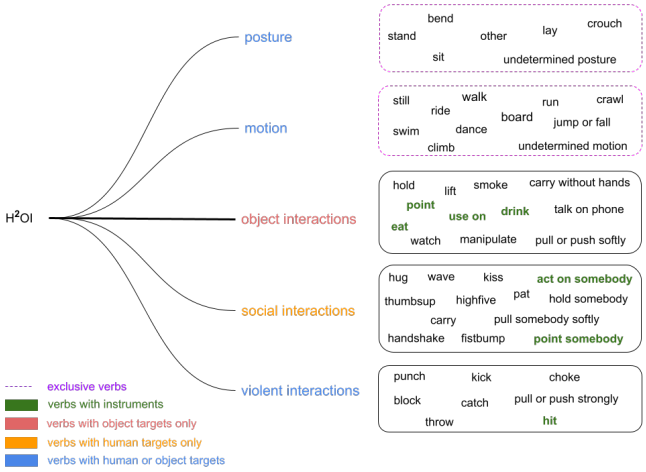

Fig. 1. H^2O taxonomy hierarchy

Posture and motion categories have verbs which are exclusive and mandatory. This means that each person in the image must be annotated with a single verb of posture and a single verb of motion. H^2O contains 58,225 posture verbs and as many motion verbs. Verbs of the other three categories are not exclusive, nor mandatory. This means that subjects can be annotated with none, one or several of these interaction verbs.

Posture verbs – General posture verbs are: "stand", "bend", "sit", "crouch", "lay", "other" and "undetermined posture". "Undetermined posture" is dedicated to truncated people whose posture cannot be determined for sure. On the contrary, "other" means the subject is fully seen but his/her posture is not usual or cannot be simply described. For example, it is the case of some acrobatic positions in sports images.

Motion verbs – Motion verbs are: "still", "walk", "run", "ride", "board", "crawl", "jump or fall", "dance", "swim", "climb" and "undetermined motion". "Still" is dedicated to people who do not move. "Undetermined motion" is annotated for people who are truncated and whose movement cannot be described for sure. We chose to annotate with

"board" all types of movement on a board (*e.g.*, skateboard, surf, snowboard, skis) not to be tied to the context of the interactions. All posture and motion verbs can have a target and it is necessarily the same for both verbs. For instance, a person can be "stand"-ing "still" on a stool. In the third row of Figure 2, the distinction between posture and motion with the possibility of annotating a target allow a precise description of a "crouching man boarding a skateboard".

Interactions with object – These verbs can be done only with objects. We finely separate the fact of holding something in four verbs: "hold", "lift", "carry without hands" and "pull or push softly" as they are visually different. "Lift" is dedicated to heavy objects which need two hands to be lifted (*e.g.*, a sofa). "Carry without hands" is used with objects which are carried without the need of hands (*e.g.*, handbag or backpack on the back or over the shoulder). "Pull or push softly" is reserved for rolling objects as suitcase, shopping cart or stroller. We do not distinguish "pull" from "push" as it is ambiguous on only one image. "Manipulate" is annotated for subjects who use an object for its specific function. For example, this verb gathers verbs as "cut", "brush" or "stick". Four verbs of this category accept until two types of interacting objects: the final target of the interaction and the tool or instrument used to execute the interaction on the target object. These verbs are: "point", "use on", "eat" and "drink". "Eat" and "drink" are annotated as such, only when the subject makes the gesture of bringing something to the mouth (*e.g.*, to sit around a table with a dish does not mean the person is eating it). Finally, "watch", "talk on phone" and "smoke" are also part of this category.

Social interactions – Verbs of this category are exclusively related to interactions between people. The interaction verbs are: "hug", "kiss", "handshake", "wave", "highfive", "fist-bump", "thumbsup", "pat", "hold somebody", "pull or push somebody softly", "carry somebody", "point somebody" and "act on somebody". "Point" and "act on" can be used with the instrument (if any) that allows the interaction achievement, as well as the final human target (*e.g.*, a doctor "acts on" a patient "with" a stethoscope).

Violent interactions – Targets of the interaction verbs of this category can be either a person or an object. Interactions are performed with strength or violence. The verbs selected highly depend on the body parts involved: "punch", "kick", "choke", "block", "pull or push strongly", "throw", "catch" and "hit". If appropriate, "hit" can be annotated with an instrument (*e.g.*, "hit" the ball "with" a baseball bat).

C. Comparison with existing datasets

V-COCO dataset [12] is a subset of COCO dataset [19] where each person is annotated with 26 interaction verbs over 80 object categories. Contrary to H^2O, four verbs ("stand", "smile", "run" and "walk") do not allow target. In addition, each interaction in restricted to only one target object for a given subject. These two points do not allow to describe the scene exhaustively. For example, if a person is standing on a stool and holding two different objects at the same time, these limitations force the partial description of a standing person holding one single object. In the image on second row of Figure 2, V-COCO annotates only one of the two suitcases pulled by the person. Moreover and contrary to H^2O, not all interactions that are part of the 26 verbs are exhaustively annotated in the image (*e.g.*, the standing still woman in the first row of Figure 2 is not annotated). Figure 2 illustrates differences between H^2O and V-COCO annotations.

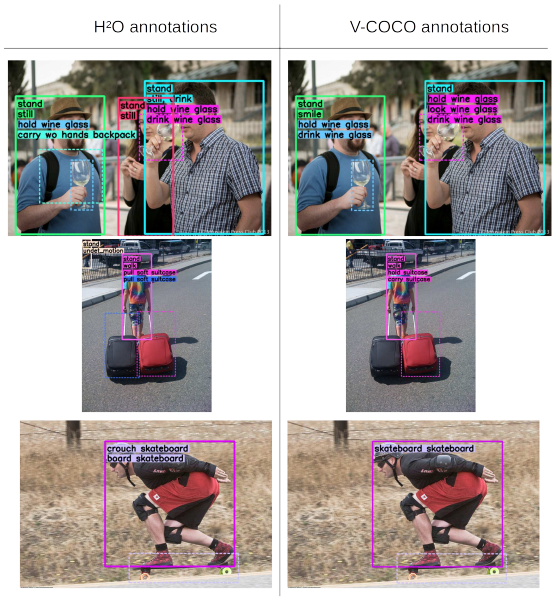

Fig. 2. Comparison between H^2O and V-COCO annotations

HICO-DET dataset [4] is larger and more diverse than V-COCO dataset because it contains 117 action categories over the same 80 object categories as COCO dataset. However, these predicates are very specific and too much linked to the context of the scene. For instance, "dribble" is always done with a sports ball or "grind" is always done with a board.

Unlike H^2O, in these two datasets, interacting objects outside the 80 classes of COCO are not annotated. Table I and II present statistics on H^2O compared to V-COCO and HICO-DET. Table I shows the amount of images, the number of verbs and their target type. Table II compares the overall number of interactions between H^2O and V-COCO, and details the numbers of interactions by categories. As posture and motion categories are mandatory, therefore exhaustively annotated in H^2O, their quantity is much larger than for V-COCO. The table also compares the mean number of people per image, and the mean number of objects per image.

TABLE I
DATASET COMPARISON ACCORDING TO NUMBER OF IMAGES AND VERBS

Dataset	#images	#verbs	Target type
HICO-DET [4]	47,774	117	object
V-COCO [12]	10,346	26	object
H^2O	13,967	51	object, person

TABLE II
DATASET COMPARISON ACCORDING TO THE AMOUNT OF ANNOTATIONS

Dataset	#interactions		#person per image	#object per image
V-COCO [12]	**Total:**	**49,019**	3.8	4.9
	Posture + Motion:	22,480		
	HOI:	26,539		
H^2O	**Total:**	**151,816**	4.2	5.1
	Posture + Motion:	116,450		
	HOI:	25,984		
	Social:	5,413		
	Violence:	3,969		

D. Evaluation protocols

V-COCO [12] proposes an evaluation of the HOI detections based on two metrics. The first one is the agent mean average precision, AP_{agent} which measures the accuracy of the pair $< subject, verb >$. The second one, the role mean average precision called AP_{role} analyzes the whole triplet $< subject, verb, target >$ considering as a true positive when all components are correct. The predicted human and object bounding boxes are supposed to be correct if they have an IoU greater than 0.5 with ground truth boxes. Two different AP_{role} metrics are proposed. They differ in the evaluation of specific interaction triplets $< subject, verb, \emptyset >$ that appear when the target object is either not seen, not existing or not in the 80 classes of COCO. In the first one (AP_{role1}), not predicting the $\emptyset$ as target is penalized whereas in the second scenario (AP_{role2}), it is not. Moreover, V-COCO dataset assumes only a single target object for a given verb and a given person. Consequently, AP_{role} computation is limited to this assumption.

H^2O dataset provides annotation of targets even though target object is not part of the 80 classes of COCO. The specific triplet $< subject, verb, \emptyset >$ in H^2O means target is not seen or not existing. Consequently, we propose a new scenario called "Objectness" where target object outside the 80 classes of COCO should be detected and properly associated to the interaction. Class label of such a target is "other".

H^2O also provides several objects in interaction for a given verb and a given person if they exist. Consequently, AP_{role} metric has been adapted to take into account this new feature. For clarity, original V-COCO scenario will be called "Original" as opposed to the new "Objectness" scenario proposed in this paper. Both dataset and evaluation code will be made available.

In this section, we present our approach for interaction detection. The method is named DIABOLO and is based on the CALIPSO [3] method. CALIPSO densely estimates interactions on a classical detection anchor grid thanks to three tasks. The first task defines the action score of each anchor. The second task estimates for each verb the presence of an interacting target for each human anchor. The third task gives an embedding for each anchor in the image in order to associate the interacting pairs. Thus, at the inference time, CALIPSO needs an external object detector to point out anchors which actually correspond to people or objects. Contrary to CALIPSO, DIABOLO integrates an object detector, based on EfficientDet [25]. Therefore DIABOLO simultaneously detects people, objects and their interactions in the image in a single shot.

Figure 3 illustrates the multi-task neural network architecture used by DIABOLO. The object detection and the interaction estimation branches share the same EfficientNet backbone followed by a BiFPN introduced by [25].

E. Object detection branch

DIABOLO uses EfficientDet [25] as object detector. EfficientDet is an anchor-based architecture as most of object detector. The outputs of the object detection branch are the classification of the instances and the regression of their bounding boxes. During training, DIABOLO supervises these two tasks in the same way as [25].

F. Interaction estimation branch

DIABOLO interaction estimation branch is close to the CALIPSO one. Indeed, this branch densely estimates three tasks on the anchor grid: verb prediction task in active and passive voice, target presence estimation task and interaction embedding task. The use of an EfficientNet backbone and BiFPN instead of a ResNet FPN allows to reduce the number of convolutions contained in the interaction network. Unlike CALIPSO, DIABOLO interaction network is composed of only two convolutional blocks after the BiFPN. The supervision of the target presence estimation task and the interaction embedding task are the same as in CALIPSO i.e. respectively a binary cross entropy loss and a metric learning pull-push loss. However, DIABOLO uses a focal loss [18] to densely predict the interaction verb. The focal loss FL aims to better cope with the class occurrence imbalance and is computed as follows:

$$FL = \sum_{v \in V} \sum_{a \in A^+} -\alpha (1 - p_a^v)^\gamma \log(p_a^v) \quad (1)$$

where v is a verb in the set V, α and γ are focal loss parameters and A^+ denotes the set of anchors associated to an interacting object. In focal loss, p being the model's estimated probability for class v and anchor a, p_a^v is equal to p if the ground truth class of anchor a is v, $1-p$ otherwise.

G. Inference

From the detection branch, people and objects are pointed out with a traditional Non Maxima Suppression (NMS) algorithm except that indices of selected anchors are used to directly read the information relative to the different interactions in the interaction branch. For each detected person, an interaction score for each verb and each possible target for this verb, according to its category (cf. Figure.1), is computed similarly to [3]. Categories of exclusive interactions are processed independently by providing only triplets of the verb with the highest probability in its category.

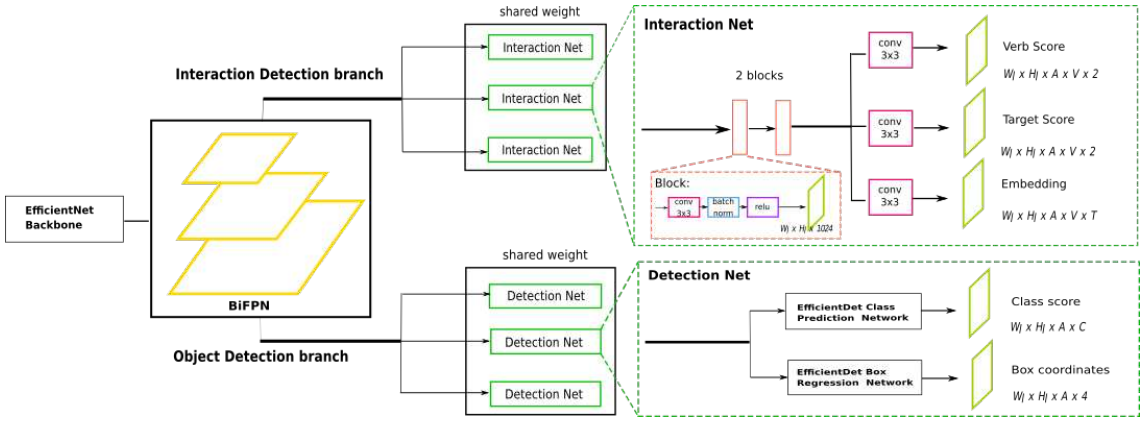

Fig. 3. DIABOLO architecture network with two branches for interaction detection (top) and object detection (bottom). W_l and H_l denote respectively the width and the height of the feature map of the pyramid at level l. A is the number of anchors, V is the number of verbs and T is the embedding size.

IV. EXPERIMENTS

In this section, we present the experiments to evaluate DIABOLO performance relative to the state of the art and propose a first baseline on H^2O dataset.

A. Datasets and Metrics

1) Datasets: We chose to evaluate DIABOLO on V-COCO [12] and on our new dataset H^2O. As mentioned in section III-C, V-COCO still has the drawback of defining predicates too closely related to the context, and HICO-DET has the same issue to a greater extent.

2) Evaluation metrics: To evaluate results on V-COCO [12], we use their standard evaluation setting, as presented in section III-D, using AP_{role1} scenario which is the most difficult one. Notice that similarly to previous work [10], [8], the verb "point" is not taken into account since it has too few samples. On H^2O, both "Original" and "Objectness" scenarios are used for evaluation.

B. Implementation details

The EfficientNet backbone, the BiFPN and the detection branch are initialized with weights previously learned on the object detection COCO dataset [19]. We use EfficientDet-D3 for fair comparison with the state of the art and EfficientDet-D1 for ablation studies as it is lighter than D3. For multi-task learning, we compare the strategies of learning instance detection branch (on COCO) and freezing it or not during the subsequent learning of interaction detection, with the strategy of continuing learning instance detection along with interaction detection. In the latter strategy, we use V-COCO or H^2O to train the whole network of DIABOLO and, at the same time, COCO (V-COCO images excluded) to continue training the detection branch only. We will see that V-COCO data are not varied enough to correctly learn instance detection. Indeed, V-COCO has only 5,400 training images whereas EfficientDet is usually learned on the 120,000 training images of COCO. That is why we use mixed batches with images from V-COCO and images from COCO to improve the object detection branch. Proportions of images from each dataset in the batch are constant during training (batch of 24 COCO + 24 H^2O images on H^2OI challenge and see Table IV for batch composition on V-COCO HOI challenge). Two different trainings are made according to evaluation scenarios mentioned in Section III-D. For the "Original" one, only target objects within 80 classes of COCO are taken into account and for the "Objectness" one, all unusual objects excluded by the 80 classes are added in a new class labeled as "other" that detection branch has to learn. Training is performed on NVIDIA A100-SXM4 GPUs. DIABOLO is trained with stochastic gradient descent (SGD), with an initial learning rate of 0.016, which is then reduced by 10 at 10,000 iterations. Focal loss parameters α and γ are respectively set to 0.25 and 2. Horizontal image flipping and color jittering are applied for data augmentation.

C. Results of DIABOLO on V-COCO and comparison with HOI state of the art

Table III presents the results of DIABOLO on V-COCO dataset compared to the best state-of-the-art methods and to one-stage methods. DIABOLO is the top-1 method outperforming current state-of-the-art method DIRV [7] by 12.1 p.p with 76.7% for AP_{agent} and 1.2 p.p with 57.3% for AP_{role}. We believe that such an increase in AP_{agent} can be explained by the subject-centric aspect of DIABOLO that is more suitable for multiple action recognition. Notice that AP_{role} result for DIABOLO is obtained without using any ontology information, contrary to [7] which loses 1.3 p.p. if no ontology information is provided.

D. Ablation studies for DIABOLO learning strategies

UnionDet [14] and DIRV [7] methods first train the object detector on COCO dataset and then freeze both backbone and instance detection branch to maintain detection performance. In this work, we choose to jointly learn detection and interaction on V-COCO and COCO datasets. To measure the impact of this strategy, we use EfficientDet-D1 backbone since training time is much faster than D3. For a fine-grained analysis of results, metrics are also computed with perfect ground truth (GT) detections, to uncorrelate instance and interaction detection performances. As shown in table

TABLE III

STATE-OF-THE-ART METHODS FOR HOI DETECTION: PERFORMANCES EVALUATED ON V-COCO TEST SET

Method	**One Stage**	$AP_{agent}(\%)$	$AP_{role}(\%)$
InteractNet [10]	✗	69.2	40.0
CALIPSO [3]	✗	-	46.4
PMFNet [27]	✗	-	52.0
ConsNet [20]	✗	-	53.2
MLCNet [24]	✗	-	55.2
UnionDet [14]	✓	-	47.5
DIRV [7] w/o prior	✓	64.7	54.8
DIRV [7] w/o flip	✓	64.1	55.2
DIRV [7]	✓	64.6	56.1
DIABOLO (ours)	✓	**76.7**	**57.3**

IV, the architecture of DIABOLO interaction branch poorly works with a frozen backbone and detection branch since the AP_{role} is only 46.2% with perfect detections. However, DIABOLO trained without freezing backbone nor detection branch, achieves 61.3% with perfect detection but only 43% with model detections. Such a result is explained by poor instance detection results (18% of mAP on V-COCO test set) of DIABOLO when learned on V-COCO only for both tasks. When keeping feeding DIABOLO with COCO along with V-COCO, instance detection performance gets back to normal (35% of mAP) and AP_{role} reaches 51.1% with model detections, showing the positive effect of joint learning.

To further assess the part of error in interaction detection induced by instance detection failure, we train DIABOLO with a better backbone, EfficientDet-D3, and once again compare results given by the model detections with those with perfect detections (cf. Table IV). DIABOLO instance detection reaches 47% of mAP on V-COCO test set. Subject-target association task is more heavily impacted by instance detection than verb estimation. Indeed, using ground truth detection only increases by 3.5 p.p AP_{agent} whereas improvement is 8.7 p.p. for AP_{role}. Additionally, interaction branch with a perfect detection reaches 80.2% for AP_{agent} but only 66.0% for AP_{role}, which leaves room for improvement on the subject-target association task.

E. Results of new baseline DIABOLO on H^2O dataset

1) Quantitative results: We evaluate DIABOLO on the new dataset H^2O. DIABOLO achieves AP_{agent} scores of respectively 41% and 40.6% for "Original" and "Objectness" scenarios. For AP_{role} metric, the scores are respectively 25.26% and 23.68% for "Original" and "Objectness" scenarios. The overall AP_{role} is lower than the one on V-COCO, suggesting that H^2O is more challenging than V-COCO. We think this is due to H^2O annotation which is more exhaustive (number of interactions per image has more than doubled), and taxonomy which is more related to the body attitude of the subject, rather than the interaction context which is no longer an extra clue. Detailed performance per verb will be made available along with H^2O dataset.

2) Qualitative results: Figure 4 shows qualitative results of DIABOLO learned on H^2O. In sample (b), (e) and (g), we can see that DIABOLO well detects HHI and their reciprocity as "hug", "punch", "highfive" and "handshake". In illustration (a), DIABOLO is able to detect that two objects are held. Similarly, in example (d), both book and umbrella are correctly detected as held. In sample (c), H^2O taxonomy allows DIABOLO to correctly detect a person standing on his/her skis. Finally, in example (f), DIABOLO correctly predicts that the suitcases are pulled and not held. This illustrates the fact that H^2O taxonomy is closer to the body attitude and the way of interacting with a target than the interaction context.

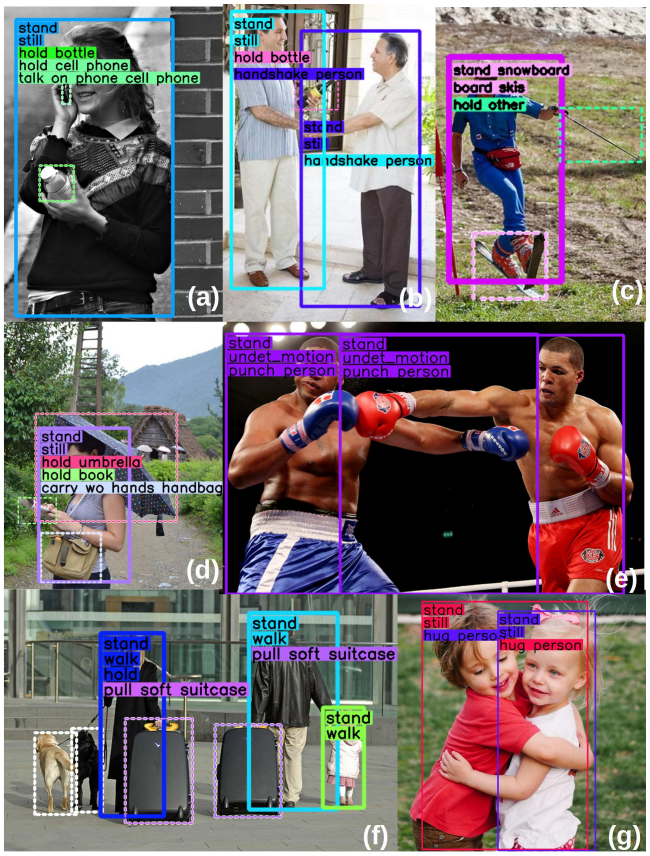

Fig. 4. Qualitative results of DIABOLO on H^2O. Interaction name has the color of the target bounding box if any, otherwise the color of the subject box. Dashed-(solid-)line box is for object (human, resp.). White boxes correspond to detected objects not interacting.

F. Computation time

Table V shows computation times of our method compared to DIRV [7]. Both methods are run on a NVIDIA RTX2080Ti GPU. DIABOLO inference time with EfficientDet-D3 is competitive with DIRV. To obtain the best performance [7] apply their method on the image and its flipped version (cf. Table III). With such a strategy, measured inference time for DIRV method is 132ms, and 129ms for DIABOLO.

V. CONCLUSION

In this paper, we propose a new challenging dataset called H^2O that copes at the same time with human interactions with other people and objects. All these interactions follow a

TABLE IV

ABLATION STUDIES FOR DIABOLO MULTI-TASK LEARNING STRATEGIES. METRICS ARE EVALUATED ON V-COCO TEST SET.

Backbone	**Detection branch**	**Batch size**	**mAP (%)**	AP_{agent} **(%)**	AP_{agent} **with GT detections (%)**	AP_{role} **(%)**	AP_{role} **with GT detections (%)**
EfficientDet-D1	frozen	12 V-COCO	42	59.8	61.9	37.3	46.2
EfficientDet-D1	learned	12 V-COCO	18	71.1	77.5	43.0	61.3
EfficientDet-D1	learned	12 V-COCO + 16 COCO	35	75.7	79.5	**51.1**	**64.0**
EfficientDet-D3	learned	24 V-COCO + 32 COCO	47	**76.7**	**80.2**	**57.3**	**66.0**

TABLE V

COMPUTATION TIME

Method	**Backbone**	**Inference Time (ms)**
DIRV [7] w/o flip	EfficientDet-D3	108
DIRV [7]	EfficientDet-D3	132
DIABOLO	EfficientDet-D1	89
DIABOLO	EfficientDet-D3	129

novel taxonomy focusing on the subject's body attitude rather than the type of the object involved or the environment. As a first baseline for H^2O, we propose DIABOLO, a new multi-task method to detect both instances and interactions without needing an external detector. This is a single-shot subject-centric method running in a fast and constant time independently of the number of instances in the image. Finally, we experimentally show that jointly training interaction and instance detections largely improves interaction detection, making DIABOLO a strong approach outperforming the state of the art on V-COCO dataset.

VI. ACKNOWLEDGMENTS

This publication was made possible by the use of the FactoryIA supercomputer, financially supported by the Ile-de-France Regional Council. Moreover, this work benefited from a government grant managed by the French National Research Agency under the future investment program with the reference ANR-19-STHP-0006.

REFERENCES

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In CVPR, pages 3686–3693, 2014.
[2] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In CVPR, pages 6299–6308, 2017.
[3] S. Chafik, A. Orcesi, R. Audigier, and B. Luvison. Classifying all interacting pairs in a single shot. In WACV, pages 2892–2901, 2020.
[4] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng. Learning to detect human-object interactions. In WACV, pages 381–389, 2018.
[5] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. HICO: A benchmark for recognizing human-object interactions in images. In ICCV, pages 1017–1025, 2015.
[6] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity recognition. In ECCV, pages 215–230, 2012.
[7] H.-S. Fang, Y. Xie, D. Shao, and C. Lu. DIRV: Dense interaction region voting for end-to-end human-object interaction detection. In The AAAI Conf. on Artificial Intelligence (AAAI), 2021.
[8] C. Gao, Y. Zou, and J.-B. Huang. iCAN: Instance-centric attention network for human-object interaction detection. In BMVC, 2018.
[9] R. Girshick. Fast R-CNN. In ICCV, pages 1440–1448, 2015.
[10] G. Gkioxari, R. Girshick, P. Dollár, and K. He. Detecting and recognizing human-object interactions. In CVPR, pages 8359–8367, 2018.
[11] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, et al. AVA: A video dataset of spatio-temporally localized atomic visual actions. In CVPR, pages 6047–6056, 2018.
[12] S. Gupta and J. Malik. Visual semantic role labeling. arXiv preprint arXiv:1505.04474, 2015.
[13] M. Hoai and A. Zisserman. Talking heads: Detecting humans and recognizing their interactions. In CVPR, pages 875–882, 2014.
[14] B. Kim, T. Choi, J. Kang, and H. J. Kim. UnionDet: Union-level detector towards real-time human-object interaction detection. In ECCV, pages 498–514, 2020.
[15] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int. Journal of Computer Vision, 123(1):32–73, 2017.
[16] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y. Wang, and C. Lu. Transferable interactiveness knowledge for human-object interaction detection. In CVPR, pages 3585–3594, 2019.
[17] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng. PPDM: Parallel point detection and matching for real-time human-object interaction detection. In CVPR, pages 482–490, 2020.
[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In ICCV, pages 2980–2988, 2017.
[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In ECCV, pages 740–755, 2014.
[20] Y. Liu, J. Yuan, and C. W. Chen. ConsNet: Learning consistency graph for zero-shot human-object interaction detection. In Int. Conf. on Multimedia, pages 4235–4243, 2020.
[21] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In ECCV, pages 852–869, 2016.
[22] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In ECCV, pages 510–526, 2016.
[23] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012.
[24] X. Sun, X. Hu, T. Ren, and G. Wu. Human object interaction detection via multi-level conditioned network. In Int. Conf. on Multimedia Retrieval, pages 26–34, 2020.
[25] M. Tan, R. Pang, and Q. V. Le. EfficientDet: Scalable and efficient object detection. In CVPR, pages 10781–10790, 2020.
[26] O. Ulutan, A. Iftekhar, and B. S. Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In CVPR, pages 13617–13626, 2020.
[27] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He. Pose-aware multi-level feature network for human object interaction detection. In ICCV, pages 9469–9478, 2019.
[28] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun. Learning human-object interaction detection using interaction points. In CVPR, pages 4116–4125, 2020.
[29] T. Zhou, W. Wang, S. Qi, H. Ling, and J. Shen. Cascaded human-object interaction recognition. In CVPR, pages 4263–4272, 2020.
[30] B. Zhuang, Q. Wu, C. Shen, I. Reid, and A. van den Hengel. HCVRD: a benchmark for large-scale human-centered visual relationship detection. In AAAI Conf. on Artificial Intelligence, volume 32, 2018.