Classifying All Interacting Pairs in a Single Shot

Sanaa Chafik[⋆†] Astrid Orcesi[⋆†] Romaric Audigier[⋆†] Bertrand Luvison[⋆†]
[⋆] CEA, LIST, Vision and Learning Lab for Scene Analysis, PC 184, F-91191 Gif-sur-Yvette, France
[†] Vision Lab, ThereSIS, Thales SIX GTS, Campus Polytechnique, Palaiseau, France
{firstname.lastname}@cea.fr

Abstract

In this paper, we introduce a novel human interaction detection approach, based on CALIPSO (Classifying ALl Interacting Pairs in a Single shOt), a classifier of human-object interactions. This new single-shot interaction classifier estimates interactions simultaneously for all human-object pairs, regardless of their number and class. State-of-the-art approaches adopt a multi-shot strategy based on a pairwise estimate of interactions for a set of human-object candidate pairs, which leads to a complexity depending, at least, on the number of interactions or, at most, on the number of candidate pairs. In contrast, the proposed method estimates the interactions on the whole image. Indeed, it simultaneously estimates all interactions between all human subjects and object targets by performing a single forward pass throughout the image. Consequently, it leads to a constant complexity and computation time independent of the number of subjects, objects or interactions in the image. In detail, interaction classification is achieved on a dense grid of anchors thanks to a joint multi-task network that learns three complementary tasks simultaneously: (i) prediction of the types of interaction, (ii) estimation of the presence of a target and (iii) learning of an embedding which maps interacting subject and target to a same representation, by using a metric learning strategy. In addition, we introduce an object-centric passive-voice verb estimation which significantly improves results. Evaluations on the two well-known Human-Object Interaction image datasets, V-COCO and HICO-DET, demonstrate the competitiveness of the proposed method (2nd place) compared to the state-of-the-art while having constant computation time regardless of the number of objects and interactions in the image.

1. Introduction

Several tasks of computer vision address the problem of understanding the semantic content of images, like visual relationship recognition. More specific than visual relationship, Human-Object Interaction (HOI) detection aims at detecting what happens and where in the image by paying exclusive attention on human-centric interactions. HOI detection is a challenging problem, essential for various applications such as activity understanding, surveillance, ambient assisted living, cobotics, etc. In the case of surveillance system, quickly understanding human-centric interactions is particularly interesting. As images may contain possibly numerous people and interactions, it is crucial for an HOI detection method to be scalable with the number of visible objects and interactions. This scalability issue motivated our work. In the following, "objects" assigned with human class are called *subjects* while those with non-human class are *targets*. More formally, HOI detection consists in determining and locating the list of triplets $< subject, verb, target >$ describing all the interactions visible in the image. Although HOI detection was classically based on video (in general, with a focus on a single action), recent approaches based on a single image have shown impressive results on detecting simultaneous interactions.

Generally speaking, image-based HOI detection task is achieved by solving the following sub-tasks: detecting interacting objects (the *object detection problem*), correctly pairing such objects (the *association problem*) and classifying the interactions (the *verb classification problem*). Most approaches [3, 8, 10, 11, 16, 24, 30, 32] rely on an object detector that identifies some candidates for subject-target pairs which boxes are then processed in a second step to assess interaction presence and type. Sometimes, features for objects and pairs are first extracted and, then, processed to infer object class, location, subject-target association and verb classification [21]. Thus, all these methods have a pair-based second-step processing, which may become a scalability issue when dealing with large numbers of object and interaction instances in the image.

This work proposes a new interaction detection approach, named CALIPSO (Classifying ALl Interacting Pairs in a Single shOt) which complexity is independent of the number of interactions. The proposed model simultaneously estimates all interactions between all objects with a

single forward pass throughout the image. It manages the problems of association and verb classification while any external object detector can be used to deal with the problem of object detection. To this end, CALIPSO approach exploits a multi-task learning scheme, performing three complementary tasks: a classification task predicts the verb of interaction, a target presence estimation task assesses the presence of the target object of the interaction and an embedding task maps a pair of interacting subject-target to a similar representation. Lastly, at inference time, any object detector can be used to point out objects of interest and output the corresponding interactions. Notice that the proposed approach does not use any ontology information such as a prior list of interactions of interest, in order to promote generalization over target classes. We have evaluated the efficiency of the proposed approach on two widely used HOI datasets. Our results compare favorably (2nd place) with state-of-the-art approaches while having constant computation time regardless of the number of objects and interactions in the image.

2. Related Work

HOI & visual relationships Despite the rapid research progress in analysis of humans and their activities by computer vision, human interaction recognition from a single image remains a challenge. Whereas videos contain rich temporal clues, such as those used in interaction analysis of egocentric videos [2, 7, 27], images contain a lot of contextual information that is meaningful to infer relationships between objects. One of the main problems of detecting visual relationships is the need for tremendous amounts of varied examples, as appearances and classes of both subject and target should vary for generalization of each interaction class. The release of large datasets [3, 11, 14, 33] has allowed the developement of several visual relationship detectors in recent years [4, 13, 15, 20, 21, 29, 30, 31, 32] as well as HOI detectors [1, 3, 8, 10, 11, 23, 24, 28].

Gupta et al. [11] successively detect a subject, classify the action and associate the target according to an interaction score. Several approaches [3, 8, 10, 16] extend an object detector model, namely Faster R-CNN [25], with extra branches either for predicting actions, estimating a probability density over the target object location for each action [10], the spatial relations of human-object pairs [3], an instance-centric attention measure [8], or filtering non-interactive human-object pairs with cross learning datasets [16]. Qi et al. [24] present a generic framework combining graphical models and deep neural network, capturing human-object interactions iteratively. Li et al. [15] introduce a cross branch communication with phrase-guided message to ensure a joint modeling of action classification and target association.

Some techniques [1, 28, 29] incorporate *linguistic knowledge* to address the issue of having a long-tail distribution of human-object interaction classes. They exploit the contextual information present in the language priors learnt with a 'word2vec' network, to generalize interactions across functionally similar objects. Alternatively, Peyre et al. [23] learn a visual relation representation combining compositional representation for subject, target and predicate with a visual phrase representation for HOI detection. Unlike these approaches, our method does not use additional linguistic data.

However, all these approaches have a pair-based processing step, i.e., a substantial processing applied on a set of subject-target pair proposals. This may become a scalability issue when dealing with large numbers of object and interaction instances in the image. In contrast, we propose a new interaction detection approach which complexity is independent of the number of interactions in the image. The model classifies interactivity on a dense sampling of all possible object locations simultaneously.

Metric learning has been applied to many different tasks, from image retrieval [6], to face recognition [26]. In addition to providing a similarity measure to compare images, it can also been used to map visual and text features to a shared feature space [5, 12] or associate features of visual elements to recognize a group of such elements. For example, Newell and Deng proposed associative embedddings to group together body joints for human pose estimation [22]. Metric learning is also applied to visual relationship detection [21, 23, 32]. In particular, Pixel2Graphs [21] produces in a single-shot manner a set of objects and interaction links represented by a graph which is deduced from two heatmaps. Then, in a second step, each of these object or connection features is passed through a fully connected network to predict interaction properties (verb, subject-target association, object class and bounding box). This second step is, thus, dependent on the number of interactions. Besides, when multiple relations are grounded in the same location, a fixed number of slots are used to manage these overlapped relations, which may be limiting for densely populated images.

CALIPSO is also based on the metric learning paradigm. But, in contrast to Pixel2Graphs, it does not use graph to explicitly model each object and each relation. Rather, it simultaneously provides associative features and interaction types for all locations of potential subjects and targets in a single shot. Another fundamental difference is that Pixel2Graphs aims to define a unique feature for each object regardless of the relation verb, and a unique feature for each relation. Differently, CALIPSO aims to define, for each interaction verb, an embedding where all objects involved in an interaction instance should have similar features. This allows a subject-target pair to have multiple interactions while solving the overlapped interaction issue. Moreover, having

a different embedding space for each verb should intuitively leave more flexibility for modeling very different types of interactions (contact interaction, distant interaction, etc.).

3. Proposed Method

In this section, we present our proposed approach, named CALIPSO, for interaction modeling. The task of human-object interaction detection consists of locating and recognizing humans and objects in a given image and identifying the actions (i.e. verbs) that connect them. Formally, locating and recognizing the set $\mathcal{T}$ of interaction triplets $< subject, verb, target >$ with $verb$ an interaction verb among V verbs. The proposed approach deals with associating and classifying subject-target interacting pairs with complexity independent of the number of interactions. To this end, CALIPSO decorrelates object detection task from the association and the interaction classification tasks. It requires an object detector only at inference time, in order to point out and classify the objects to be really considered for interaction. We first give an overview of the proposed approach, then detail the proposed model tasks. Last, we describe the inference process.

3.1. Overview

The proposed model architecture is a multi-task neural network (cf. Figure 1). It consists of a backbone network and an interaction network. From an image I, a feature pyramid is constructed using a Feature Pyramid Network (FPN) [17] backbone, capturing multi-scale high-level semantics. The FPN backbone network takes an input image I of size $W \times H$, and outputs multiple level feature maps F_l of size $W_l \times H_l$, where $W_l = \frac{W}{2^l}$, $H_l = \frac{H}{2^l}$ and l is the pyramid level, $l \in [l_{min}, l_{max}]$. The FPN is built on top of the Residual Network following RetinaNet [18] architecture.

Then, the featurized image from each FPN level feeds a fully convolutional interaction network ending with three tasks. The first task is an action classification that predicts the verb describing the type of interaction between subject and target. The second task is a target presence estimator providing the probability that the object, a human is interacting with, is visible or not, for a given verb. The third task associates interacting subject and target, by mapping them to the same representation. The overall network is trained end-to-end, the three tasks are trained simultaneously, sharing the common backbone which subsequently helps generalization by regularizing training. CALIPSO approach simultaneously estimates all possible interactions between all humans and objects in the image, with a single forward pass through the architecture. Thus, CALIPSO is independent of the number of subjects, targets and interaction instances. Moreover, by densely estimating embeddings for each verb, negative example mining is exhaustive over the image. For example, all people not doing a specific action over the image will be provided to the network as negative samples to learn the embedding space of this action.

Finally, at inference, after generating dense maps, an external object detector is used to point out candidate subjects and targets. Therefore, the final interaction triplets are determined thanks to the object class of targets provided by the detector together with the association information and interaction verb given by CALIPSO.

3.2. Interaction module

Firstly, interaction detection requires to identify the (human and non-human) objects in interaction. At each feature map location, a set of reference boxes called anchors are defined. These anchors are of multiple scales and aspect ratios aligned to objects. We use anchor boxes similar to those in the Region Proposal Network of [9]. For each level l of the feature pyramid, we define a set of anchors $\mathcal{A}_l$, containing $W_l \times H_l \times A$ anchors, where $A = 9$ is the number of anchors at each feature map location. For sake of clarity, we define $\mathcal{A} = \{a_i | \, i \in [1, A_{all}]\}$ as the set of all the anchors over the pyramid, where A_{all} is the total number of anchors. Each anchor in $\mathcal{A}$ is labeled as foreground or background. We denote $\mathcal{G} = \{g_j | \, j \in [1, B]\}$ as the set of ground-truth bounding boxes, where B is the number of objects in the image. As it is classically done [18], an anchor is assigned to a ground-truth box if its intersection-over-union (IoU) is over 0.5. We define $\mathcal{A}_{g_j}$ as the set of anchors assigned to ground-truth boxes g_j and $\mathcal{A}_\mathcal{G}$ the union of all the anchors assigned to a ground-truth box.

The interaction subnet is responsible for three tasks learnt simultaneously. This subnet is applied with the same weights to each level of the backbone feature pyramid, capturing the relationships between instances of different sizes that occurred in different levels of the FPN. Moreover, the shared weights of the network applied to each pyramid level enhance the learning of correlated tasks. These tasks share a succession of eight blocks of convolution, batchnorm and ReLU layers. The number of blocks was found empirically (cf. section 4.5.1). The spatial size of each task output for a given pyramid level l is equal to the feature map size at this level: $W_l \times H_l$.

Verb prediction task: Considering that the subjects of the interaction can take simultaneously multiple actions, the verb prediction task minimizes a multi-label binary cross entropy loss L^{verb} between the predicted and the ground-truth verbs. Unlike other methods, we introduce an additional object-centric passive verb estimation to reciprocally improve the relationship detection. The verb prediction task is performed based on the contextual appearance which is very informative to distinguish actions that humans carry out and objects undergo. Among $\mathcal{A}_\mathcal{G}$, we find the active anchors, representing anchors associated

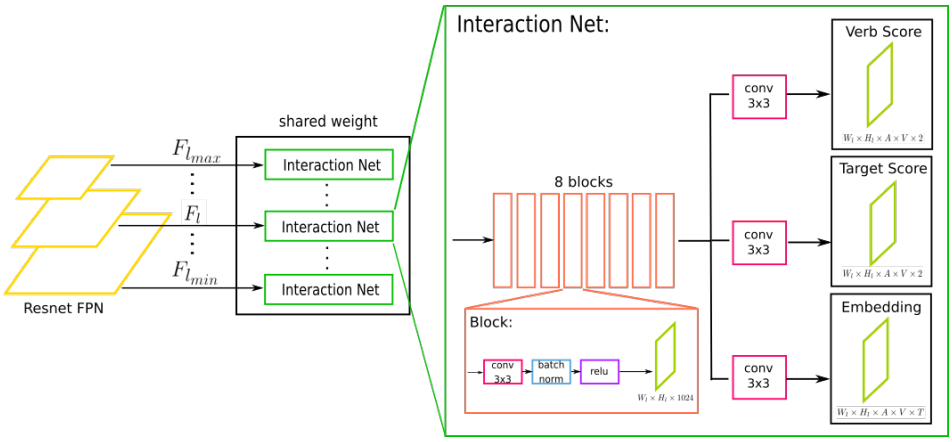

Figure 1. CALIPSO architecture starts with a Resnet FPN backbone with feature pyramid ($F_{l_{min}}$ to $F_{l_{max}}$). The feature of a given level l has size $W_l \times H_l$. The interaction network is applied on each level. It is composed of a succession of 8 blocks. The network splits into 3 branches computing 3 complementary tasks. A is the number of anchors, V is the number of verbs and T is the size of the embedding.

to people executing the actions, and the passive anchors associated to objects undergoing the action. The passive form classification is an optional task that improves performances (cf. Section 4.5.1). Verb prediction task then produces, for each anchor, a classification output over the verbs in both active and passive forms resulting in an output of size $2V$ with V the number of different verbs. This reciprocal interaction estimation is expected to be a soft way to enforce the interaction verb classification by merging human-centric and object-centric information.

Target presence estimation task is a complementary task to the verb prediction task. It aims to estimate the probability that the object, a person is interacting with, is visible or not. Similar to the verb prediction task, the target object estimation is performed on the contextual appearance of each person anchor, capturing the spatial position and the surroundings of the person in the image. For each anchor, the output of size $2V$ consists of binary sigmoid classifiers. The objective of the training is to minimize the binary cross entropy loss, L^{target}, between the ground-truth target object labels and predicted target estimation.

The interaction embedding task aims to map several anchors corresponding to interacting subject and target to the same representation for a given verb. The embedding subnet is a function mapping the anchor space $\mathcal{A}$ to a new space such that: $emb\colon \mathcal{A} \rightarrow \mathbb{R}^{V \times T}$ where T is the dimension of the interaction embedding space specific to one verb. For a given verb, this embedding task aims at ensuring to assign, first, the same embedding to anchors related to the same object instance and, second, the same embedding to anchors belonging to the same interaction.

Formally, given anchors $a_i, a_j \in \mathcal{A}^2_{\mathcal{G}}$, a_i and a_j are interacting according to verb v, i.e. $a_i \sim_v a_j$, if:

$$\exists\, g_n \in \mathcal{G} \mid (a_i, a_j) \in \mathcal{A}^2_{g_n} \tag{1}$$

or

$$\exists\, (g_n, g_m) \in \mathcal{G}^2,\ n \neq m, \\ < g_n, v, g_m > \text{or} < g_m, v, g_n > \in \mathcal{T} \Big| (a_i, a_j) \in \mathcal{A}_{g_n} \times \mathcal{A}_{g_m} \tag{2}$$

Accordingly, to each verb v, corresponds a set of equivalence classes associated with an equivalence relation $\sim_v$, denoted by $\mathcal{C}_v = \{\mathcal{C}^i_v | i \in [1, E_v]\}$, with E_v the number of equivalence classes for verb v. Let $|\mathcal{C}^i_v|$ be the number of anchors belonging to the equivalence class $\mathcal{C}^i_v$. The reference of the equivalence class is defined by the mean of the output embeddings of the same equivalence class as follows:

$$\overline{e_{\mathcal{C}^i_v}} = \frac{1}{|\mathcal{C}^i_v|} \sum_{j \in \mathcal{C}^i_v} e^v_j \tag{3}$$

where e^v_j is the predicted embedding for the anchor a_j and verb v.

The embedding network aims to learn the equivalence class space $\mathcal{C}_v$, by minimizing the equivalence loss L^{emb}_v, defined in a metric learning form. For a given verb v, the loss is defined as:

$$L^{emb}_v = L^{pull}_v + L^{push}_v \tag{4}$$

The pulling loss that aims at gathering the corresponding elements, is defined as:

$$L^{pull}_v = \frac{1}{E_v} \sum_{\mathcal{C}^i_v \in \mathcal{C}_v} \frac{\lambda_{\mathcal{C}^i_v}}{|\mathcal{C}^i_v|} \sum_{j \in \mathcal{C}^i_v} \left(e^v_j - \overline{e_{C^i_v}}\right)^2 \tag{5}$$

Based on the ground truth annotations defining interacting instances, the first term of the equation aims to merge

interacting instances to the same equivalence class by computing the mean squared distance between the equivalence references $\overline{e_{C_v^i}}$ and the predicted embedding e_j^v for each anchor j in equivalence class $\mathcal{C}_v^i$. The weight $\lambda_{\mathcal{C}_v^i}$ aims at focusing more on equivalence classes representing real interacting subjects and targets rather than equivalence class associated to a single object not belonging to any interaction (cf. equation 1). It is defined as:

$$\lambda_{\mathcal{C}_v^i} = \begin{cases} \lambda_{pull} & \text{if } \exists\, a_j, a_k \in \mathcal{C}_v^i \text{ such that} \\ & (a_j, a_k) \in \mathcal{A}_{g_n} \times \mathcal{A}_{g_m}, n \neq m, \\ & < g_n, v, g_m > \text{or} < g_m, v, g_n > \in \mathcal{T}; \\ 1 & \text{otherwise.} \end{cases} \tag{6}$$

The pushing loss enables the mapping of not interacting instance anchors into different clusters using an exponential decreasing function with fixed parameter σ. It is defined as:

$$L_v^{push} = \frac{1}{E_v^2} \sum_{\substack{\mathcal{C}_v^i, \mathcal{C}_v^j \in C_v^2 \\ i \neq j}} \gamma_{\mathcal{C}_v^i, \mathcal{C}_v^j}\, exp\left(\frac{-1}{2\sigma^2}\left(\overline{e_{\mathcal{C}_v^i}} - \overline{e_{\mathcal{C}_v^j}}\right)^2\right) \tag{7}$$

The weight $\gamma_{\mathcal{C}_v^i, \mathcal{C}_v^j}$ introduces a soft penalty to the loss to force the network to associate the correct target among several objects present in the image that are usual target for this verb. For example, the feature of a person sitting on a given chair should not be clustered with features of other chairs or objects one can sit on (e.g. couch, bed, table, ...), present in the image. This pushing weight is a way to enforce the selection of the right target among various candidates even if they are suitable for this interaction. More formally, let lab_i be the class label of anchor a_i and $\mathcal{L}_v$ the set of object classes that can be involved in the type of interaction given by verb v according to statistics on the dataset (e.g. chair, couch, bed, table, ... for verb ''sit"). The weight $\gamma_{\mathcal{C}_v^i, \mathcal{C}_v^j}$ is defined as :

$$\gamma_{\mathcal{C}_v^i, \mathcal{C}_v^j} = \begin{cases} \gamma_{push} & \text{if } \exists (a_k, a_l) \in \mathcal{C}_v^i \times \mathcal{C}_v^j \text{ such that} \\ & (a_k, a_l) \in \mathcal{A}_{g_n} \times \mathcal{A}_{g_m},\ n \neq m, \\ & (lab_k, lab_l) \in \mathcal{L}_v^2; \\ 1 & \text{otherwise.} \end{cases} \tag{8}$$

This embedding scheme is performed for each verb, allowing the network to learn the different ways of interaction depending on the verb. Moreover, the embedding predictions are performed simultaneously on all anchors, regardless of the number of object instances. This also enables a better management of negative interactions at training by processing all non-interactions in the image. In addition, it allows a fast and accurate instance connection at inference. Notice that the embedding task does not make specific assumptions between subject and target positions and can thus model both distant and close interactions. In addition, the embedding task learns to associate objects of possibly different sizes, i.e., localized on different pyramid levels.

The overall loss L_{total} of the proposed model is the sum of verb classification loss L^{verb}, the target presence loss L^{target}, and the mean of embedding losses L_v^{emb}.

$$L_{total} = L^{verb} + L^{target} + \frac{1}{|V|} \sum_{v \in V} L_v^{emb} \tag{9}$$

3.3. Inference

In the same manner as existing approaches, we predict the HOI triplets $< subject, verb, target >$, which involves predicting the human-object bounding box pairs, identifying the verb and the triplet score. The three tasks of the proposed model provide three feature anchor maps. The feature anchor map of the first task defines the action score of each location in the image. The second task provides a feature map estimating for each verb, the presence of an interacting target for each human anchor. The third feature anchor map gives an embedding for each anchor in the image, to determine the interacting anchors. The method extracts all the feature maps simultaneously and independently of the number of object instances which are at arbitrary image locations and scales, contrary to most existing approaches where every selected human-object pair is processed individually.

The prediction of HOI triplets requires preliminary to identify all human-object bounding boxes. For that purpose, CALIPSO requires at inference an external detector to point out anchors of interest from the three feature maps. The external detector can be any bounding box-based object detector providing the bounding box positions and the class scores, noted s_h^{det} for human and s_o^{det} for object. The detector provides a set of candidate object bounding boxes that are subsequently mapped to the anchor grid. Hence, from this mapping, for each verb v and for each candidate bounding box, different scores can be read: verb scores (specifically, active score $s_{v,h}^{active}$ for human and passive score $s_{v,o}^{passive}$ for object), target presence scores $s_{v,h}^{target}$ for human, and embeddings e_i^v of each detected instance. These embeddings are compared each other defining a connection score $s_{v,h,o}^{emb}$ computed as follows:

$$s_{v,h,o}^{emb} = exp^{(-|e_h^v - e_o^v|)} \tag{10}$$

All the above scores together define the triplet score as:

$$s_{v,h,o}^{triplet} = \sqrt[6]{s_h^{det}\ s_{v,h}^{active}\ s_o^{det}\ s_{v,o}^{passive}\ s_{v,h}^{target}\ s_{v,h,o}^{emb}} \tag{11}$$

All the possible triplets are computed for each detected human and each verb. Additionally, a pair score is computed for target absence case:

$$s_{v,h}^{pair} = \sqrt[3]{s_h^{det}\ s_{v,h}^{active}\ (1 - s_{v,h}^{target})} \tag{12}$$

For a given verb and a given person, all triplets and the pair are sorted according to their scores and the one with the highest score is kept after thresholding.

4. Experiments

Experiments are conducted on two widely used datasets for interaction detection with a comparison between the proposed approach and recent state-of-the-art.

4.1. Datasets

V-COCO dataset[1] [11] is a subset of the COCO dataset [19] for human-object interaction detection. It includes $10,346$ images ($2,533$ images in the train set, $2,867$ images in the validation set and $4,946$ images in the test set). V-COCO contains $16,199$ human instances, where each person has annotations for 29 action categories over 80 object categories. The target objects of the dataset are classified into two types: "object" or "instrument": "object" target if it undergoes the action (e.g., "to cut a cake"), or "instrument" if it is a means enabling the interaction (e.g., "to cut with a knife"). Four verbs do not have target ("stand", "smile", "run", "walk")

HICO-DET dataset [3] is a subset of the HICO dataset for human-object interaction detection. It is larger and more diverse than V-COCO dataset. HICO-DET includes $47,051$ images ($37,536$ images in the train set and $9,515$ images in the test set). HICO-DET contains 117 action categories over 80 object categories as COCO dataset. Not all combinations of actions over objects are relevant, according to a defined ontology. As a consequence, only 600 specific human-object interaction categories are annotated and evaluated.

4.2. Evaluation metrics

Following the standard evaluation settings of V-COCO [11] and HICO-DET [3] datasets, we evaluate HOI detection performance using the average precision metrics. The predicted $< subject, verb, target >$ triplet is considered as a true positive, when all the triplet predicted components are correct. The predicted human and object bounding boxes are supposed to be correct if they have IoU greater than 0.5 with ground truth boxes.

Following previous work [3, 8, 10, 24], the evaluation on V-COCO dataset is based on the role mean average precision called AP^1_{role} on 24 verb categories. Indeed, for the purpose of fair comparison with state-of-the-art approaches, 5 actions (run, smile, stand, walk and point) are ignored in the evaluation, as done in previous approaches.

Concerning HICO-DET dataset [3], we report the mean AP over three different HOI category sets: (a) all 600 HOI categories in HICO (*Full*), (b) 138 HOI categories with less than 10 training instances (*Rare*), and (c) 462 HOI categories with 10 or more training instances (*Non-Rare*).

4.3. Implementation details

We initialize the FPN ResNet backbone with corresponding weights of RetinaNet [18] especially trained on COCO dataset from which V-COCO images were previously removed. The CALIPSO is trained with stochastic gradient descent (SGD), with an initial learning rate of 0.016, which is then reduced by 10 at 25000 iterations over a batch of size 10. A horizontal image flipping is applied for data augmentation. The weight decay is set to 10^{-4} and the momentum to 0.9. σ, λ_v and γ_v are experimentally set to 2, 10 and 100.

At inference, CALIPSO requires an external detector to filter interacting bounding boxes from the three sub-task feature maps. As done in most state-of-the art methods, the Faster RCNN [25] from Detectron[2] framework is used as external detector. It is based on a ResNet-50-FPN backbone to generate all object bounding boxes. Other object detectors are tested to show the influence on HOI detection.

4.4. Qualitative results

Figures 2 and 3 illustrate the interaction results detected by the proposed model. They show all the triplets occurred in the image. Each triplet is represented by a solid-line box for the subject and a dashed-line box for the target object. At the top left of the subject box, the action performed is indicated on a background with same color as the related target box.

Figure 2 depicts interactions detected by our approach. As can be seen, CALIPSO can infer HOI in various situations such as: 1) Individual person performing different actions on a single object (i.e, "a person rides, sits and holds a bicycle", in Figure 2-a-b-c-f). 2) Individual person interacting with different objects (e.g., in Figures 2-b and 2-f, "a person works on a computer while sitting on a chair/couch"). 3) Several people interacting with a single object (e.g., in Figure 2-e, "two people hold the same knife"). Notice that CALIPSO correctly assigns the target object to the corresponding action, and can successfully detect contactless interactions (in Figure 2-d, "look at and throw a frisbee").

Figure 3 illustrates another sample of V-COCO test images, where CALIPSO detects some incorrect triplets. This is mainly caused by : 1) Wrong object detection, with either no detected object (as shown in Figure 3-b where the cell phone is not detected) or misclassified object (illustrated in Figure 3-c where the backpack is classified as human). 2) Wrong verb estimation, depicted in Figure 3-d where the person has a confusing posture. 3) Wrong target association, shown in Figure 3-c where the wine glass is held by

[1] https://github.com/s-gupta/v-coco

[2] https://github.com/facebookresearch/Detectron

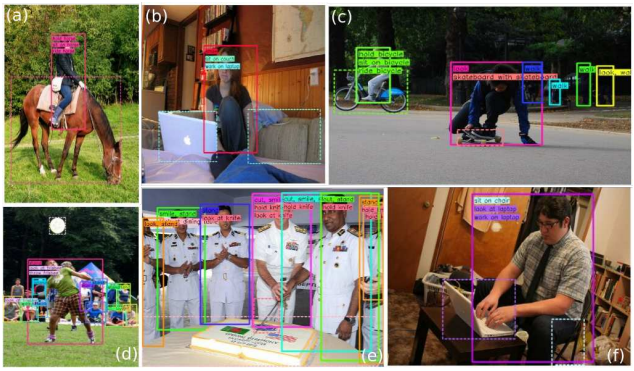

Figure 2. Samples of human-object interactions detected by CALIPSO on some V-COCO test images. An interaction triplet is composed of a human subject represented by a solid-line box, a target object represented by a dashed-line box and, at the top left of the subject box, the action performed is written on a background with same color as the target object box (best viewed in color).

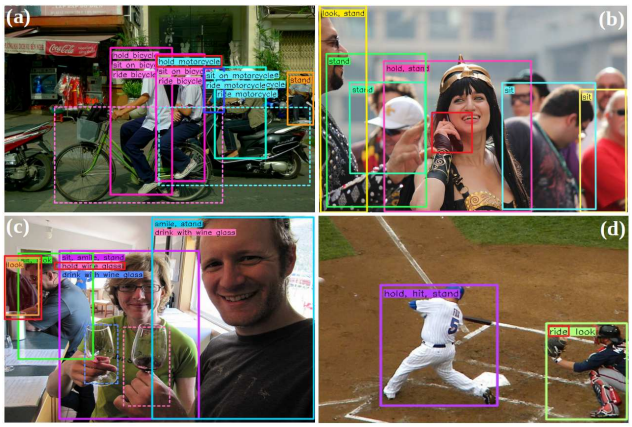

Figure 3. Illustration of some incorrect human-object interaction detections on some V-COCO test images.

the wrong person. Figure 3-a shows an example where all these difficulties appear simultaneously. Indeed, the high density of objects leads to more occlusion, misunderstanding of the object depth in the scene and, thus, confusions in subject-target associations.

4.5. Quantitative results

4.5.1 Ablation study

In Table 1 we evaluate on V-COCO dataset the contributions of various components of the method.

Shared weights: Sharing weights across feature pyramid network levels shows improvement by 2.25 p.p. (percentage points) in interaction detection performances. Intuitively, it may capture better the relationships between instances belonging to different levels of the FPN corresponding to different object size.

Passive mode: Whereas active mode is subject-centric, the passive mode is a way of introducing a complementary target-centric point of view and, thus, introducing redundancy to improve robustness. Without passive mode task, our model reaches an AP^1_{role} of 36.86%. It increases by approximately almost 10 p.p. and reaches an AP^1_{role} of 46.36% when passive mode task is used.

Method	$\mathbf{AP^1_{role}}$ (%)
CALIPSO	**46.36**
CALIPSO w/o weight sharing	43.86
CALIPSO w/o passive mode	36.86
CALIPSO w/o target presence	25.51
CALIPSO 5 blocks	44.35
CALIPSO 8 blocks	**46.36**
CALIPSO 11 blocks	45.05

Table 1. Ablation studies for CALIPSO on theV-COCO test set.

Target presence: Target presence has on CALIPSO performance a huge impact, increasing results by about 20 p.p. Such a variation in performance is due to the difficulty of setting a maximal distance (in the embedding) below which a subject can be considered in interaction with the target. It is well-known that directly thresholding a learnt metric is not trivial. Indeed, metric learning does not constrain absolute distance between samples but only a ranking between them. Target presence task is a way to bypass this issue.

Depth of Interaction Net: The number of blocks used in the Interaction Net has been empirically chosen. A succesion of 8 blocks showed the best result.

4.5.2 Results on V-COCO dataset

As the proposed method focuses on HOI classification independently of object detection task, it can advantageously use any external object detector at inference time. Indeed, changing the detector does not require to re-train or adapt the network, which is a very interesting property when better object detectors appear in the state of the art. Consequently, we evaluate our model with two different external object detectors in input: Faster RCNN [25] with ResNet50 backbone (*Faster R50*) which is generally used by state-of-the-art methods as a basis to learn interactions, and Faster RCNN with a ResNext101 backbone (*Faster RNext101*). For fair comparison, we report RP_DC_D results of Interactiveness [16] approach which corresponds to models trained without extra datasets.

Table 2 shows the evaluation results of CALIPSO variants compared to state-of-the art methods on V-COCO dataset. CALIPSO reaches the second place behind Interactiveness [16] but it is computationally far more efficient as we will see in Section 4.5.4.

Besides, in order to decorrelate object detection task from interaction detection one, we use at inference the perfect object detector and report results in table 2. The per-

formance is increased by about 7 p.p. which shows that the main issue is still the interaction detection (i.e. verb classification and subject-target association).

Method	Detector / BB	AP^1_{role} (%)
VSRL [11]	Faster R50	31.8
InteractNet [10]	Faster R50	40.0
GPNN [24]	Deform. CNN	44.0
iCAN late(early) [8]	Faster R50	44.7 (45.3)
Xu [28]	Faster R50	45.9
Interactiveness [16]	**Faster R50**	**47.8**
Ours	Faster R50	46.36
Ours	Faster RNext101	47.65
Ours	Groundtruth	54.48

Table 2. Evaluation results for CALIPSO on V-COCO test set compared with state-of-the-art methods. Object detectors or backbones (BB) used are mentioned in the middle column.

4.5.3 Results on HICO-DET dataset

Since objects in HICO-DET dataset are loosely annotated (many boxes can be assigned to the same object), we adopt the same protocol as [10] to clean annotation. We use a ResNext101 object detector trained on COCO to detect object and assign the ground truth labels from HICO-DET annotations to the detected objects that highly overlap HICO-DET boxes.

Following the evaluation settings of [3], we report the quantitative evaluation of *Full*, *Rare*, and *Non-Rare* interactions on "*default*" evaluation setting. Table 3 reports the average precision results of our method on HICO-DET dataset, compared to state-of-the-art HOI detection approaches. Once again, for fair comparison, we reported methods that only use the dataset without help of additional data, such as linguistic knowledge, from external datasets.

The proposed approach shows competitive results reaching the second place with Faster RNext101 detector.

	Average Precision (*Default*)		
Method	**Full**	**Rare**	**Non-Rare**
HO-RCNN [3]	7.81	5.37	8.54
InteractNet [10]	9.94	7.16	10.77
GPNN [24]	13.11	9.34	14.29
Xu [28]	14.70	13.26	15.13
iCAN [8]	14.84	10.45	16.15
Interactiveness [16]	**17.03**	**13.42**	**18.11**
Ours (Faster R50)	14.31	10.43	15.46
Ours (Faster RNext101)	14.89	11.12	16.01

Table 3. Evaluation results on HICO-DET test set compared with state-of-the-art methods.

4.5.4 Computation Complexity and Time

Concerning complexity relative to the numbers of people (N) and objects (M) in the image, notice that CALIPSO only does one pass throughout the image with complexity $O(1)$, whereas all other state-of-the-art approaches have a complexity of $O(P)$ with P the number of processed pairs, $T \leq P \leq N \times M$ with $T = |\mathcal{T}|$ the number of ground truth triplets. The impact on computation time is shown in Figure 4: CALIPSO runs in constant time (460 ms on NVIDIA Titan X Pascal) independently of the numbers of people and objects in the image. Differently, state-of-the-art methods which provide their codes, Interactiveness [16] and iCAN [8], have a soaring computation time (e.g., from less than 1 second to more than 40 seconds for Interactiveness).

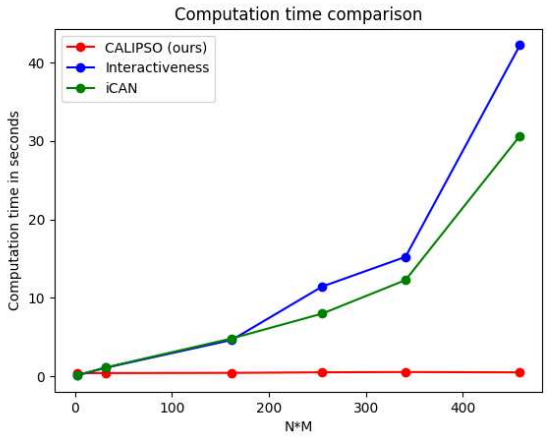

Figure 4. Computation time in seconds for CALIPSO (ours), Interactiveness [16] and iCAN [8] for increasing numbers of potential pairs present in images.

5. Conclusion

In this paper, we proposed a novel interaction detection model, named CALIPSO. It estimates all interactions efficiently and simultaneously between all human subjects and object targets by performing a single forward pass throughout the image, regardless of the numbers of objects and interactions in the image. This constant complexity is achieved thanks to a metric learning strategy that clusters subject and target in interaction, and pushes away all non-interacting objects. Besides, adding a target presence estimation task as well as an object-centric passive-voice verb estimation for redundancy showed performance improvement. The proposed method shows competitive results on two widely used datasets, compared to the state of the art, while being much more scalable with the number of interactions in the image.

This work was partly supported by Conseil régional d'Ile-de-France. Training was performed on Factory-IA, CEA computer facilities.

References

[1] A. Bansal, S. S. Rambhatla, A. Shrivastava, and R. Chellappa. Detecting human-object interactions via functional generalization. *arXiv preprint arXiv:1904.03181*, 2019.

[2] F. Baradel, N. Neverova, C. Wolf, J. Mille, and G. Mori. Object level visual reasoning in videos. *Proceedings of the IEEE European Conference on Computer Vision*, 2018.

[3] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng. Learning to detect human-object interactions. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2018.

[4] B. Dai, Y. Zhang, and D. Lin. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, pages 3298–3308. IEEE, 2017.

[5] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems 26*, pages 2121–2129, 2013.

[6] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[7] A. Furnari, S. Battiato, K. Grauman, and G. M. Farinella. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 2019.

[8] C. Gao, Y. Zou, and J.-B. Huang. iCAN: Instance-centric attention network for human-object interaction detection. In *British Machine Vision Conference*, 2018.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[10] G. Gkioxari, R. Girshick, P. Dollár, and K. He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367. IEEE, 2018.

[11] S. Gupta and J. Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

[12] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676, April 2017.

[13] A. Kolesnikov, C. H. Lampert, and V. Ferrari. Detecting visual relationships using box attention. *arXiv preprint arXiv:1807.02136*, 2018.

[14] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[15] Y. Li, W. Ouyang, and X. Wang. ViP-CNN: A visual phrase reasoning convolutional neural network for visual relationship detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[16] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y.-F. Wang, and C. Lu. Transferable interactiveness knowledge for human-object interaction detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[17] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 4, 2017.

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[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 *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[20] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *Proceedings of the European Conference on Computer Vision*, pages 852–869. Springer, 2016.

[21] A. Newell and J. Deng. Pixels to graphs by associative embedding. In *Advances in neural information processing systems*, pages 2171–2180, 2017.

[22] A. Newell, Z. Huang, and J. Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems 30*, pages 2277–2287, 2017.

[23] J. Peyre, I. Laptev, C. Schmid, and J. Sivic. Detecting rare visual relations using analogies. *arXiv preprint arXiv:1812.05736*, 2018.

[24] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision*, pages 407–423. Springer, 2018.

[25] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[26] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. 2015.

[27] S. Sudhakaran, S. Escalera, and O. Lanz. LSTA: Long short-term attention for egocentric action recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[28] B. Xu, Y. Wong, J. Li, Q. Zhao, and M. S. Kankanhalli. Learning to detect human-object interactions with knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[29] R. Yu, A. Li, V. I. Morariu, and L. S. Davis. Visual relationship detection with internal and external linguistic knowledge distillation. *Proceedings of the IEEE international conference on computer vision*, 2017.

[30] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi. Neural motifs: Scene graph parsing with global context. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[31] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, volume 1, page 5, 2017.

[32] J. Zhang, K. J. Shih, A. Elgammal, A. Tao, and B. Catanzaro. Graphical contrastive losses for scene graph generation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[33] B. Zhuang, Q. Wu, C. Shen, I. Reid, and A. v. d. Hengel. Care about you: towards large-scale human-centric visual relationship detection. *arXiv preprint arXiv:1705.09892*, 2017.