Interaction Compass: Multi-Label Zero-Shot Learning of Human-Object Interactions via Spatial Relations

Dat Huynh
Northeastern University
huynh.dat@northeastern.edu

Ehsan Elhamifar
Northeastern University
e.elhamifar@northeastern.edu

Abstract

We study the problem of multi-label zero-shot recognition in which labels are in the form of human-object interactions (combinations of actions on objects), each image may contain multiple interactions and some interactions do not have training images. We propose a novel compositional learning framework that decouples interaction labels into separate action and object scores that incorporate the spatial compatibility between the two components. We combine these scores to efficiently recognize seen and unseen interactions. However, learning action-object spatial relations, in principle, requires bounding-box annotations, which are costly to gather. Moreover, it is not clear how to generalize spatial relations to unseen interactions. We address these challenges by developing a cross-attention mechanism that localizes objects from action locations and vice versa by predicting displacements between them, referred to as relational directions. During training, we estimate the relational directions as ones maximizing the scores of ground-truth interactions that guide predictions toward compatible action-object regions. By extensive experiments, we show the effectiveness of our framework, where we improve the state of the art by 2.6% mAP score and 5.8% recall score on HICO and Visual Genome datasets, respectively.[1]

1. Introduction

Multi-label learning is the important yet challenging task of recognizing all labels in an image with applications in human-computer interaction, robotics, assistive technologies and surveillance systems. Due to the high cost of collecting training samples for all possible labels, multi-label zero-shot learning aims to recognize unseen labels that do not have training images [1, 2, 3]. However, the majority of existing works have focused on the case where each label is a simple concept (e.g., an object) and have tired to capture *inter-label dependencies* (e.g., co-occurrences of objects) for more effective recognition. On the other hand, richer representation and description of images require more complex labels. Human-object interactions are one such important form of labels, where each label describes an action performed on an object (e.g., 'holding cup' or 'removing wheel') [4, 5, 6, 7, 8]. However, existing multi-label learning works ignore *intra-label dependency*, which is the spatial relation between the action and the object within an interaction label. This leads to a lack of ability to distinguish between objects in backgrounds and in interactions, see Figure 1 and poses challenges to generalization to unseen interactions. The work in [9] made the first attempt in generalizing multi-label learning to human-object interaction (HOI) recognition in the zero-shot setting, where some interactions do not have training images. Our paper makes advances on this task by capturing spatial relations among actions and objects (intra-label dependencies) to enhance seen and unseen interaction recognition without requiring bounding-box of locations of actions and objects in images.

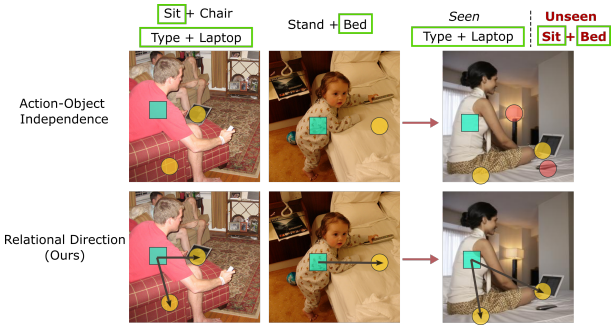

Figure 1: Conventional multi-label zero-shot recognition (top) assumes independence between action and object components in interaction labels, hence, cannot distinguish between objects in background (red) and in interactions (yellow). Our method (bottom) leverages relational directions to guide predictions toward compatible objects and actions.

Prior Works and Challenges. Most multi-label learning works take advantage of label correlations to regularize predictions [10, 11, 12, 13]. To further enhance the performance, [14, 15, 16] use attention mechanisms to extract discriminative visual features of labels. While [17, 18, 19] predict attention regions recurrently, [20, 21, 22, 7] propose

[1]Code is available at https://github.com/hbdat/iccv21_relational_direction.

specialized attention modules for object and action labels. However, these works assume that every label has training samples, therefore cannot generalize to unseen labels.

A few works have addressed multi-label zero-shot recognition by exploiting semantic information overlap between seen and unseen labels. [23] proposes a nonlinear embedding between visual features and label semantics, while [24, 2] employ external knowledge in the form of label graphs. Recently, [3] has proposed a shared attention mechanism among labels to effectively learn unseen labels. While these methods can be extended to address recognition of interaction labels instead of simple labels, they do not capture spatial dependencies between actions and objects in interactions, which as we show leads to low performance.

To handle unseen interactions, [9] proposed to capture image context for HOI prediction. However, the approach ignores discriminative spatial relations between actions and objects, necessary to determine whether objects are being interacted with by actions. Thus, recent works [25, 26, 27, 28, 29] detect human and object bounding boxes to determine their spatial compatibility. However, these works require expensive bounding-box supervision of humans and objects from seen interaction labels and are difficult to scale to thousands of interaction labels.

Paper Contributions. To address the above limitations, we propose a compositional multi-label zero-shot interaction learning framework that incorporates action-object spatial dependencies in interactions and does not require expensive bounding-box annotations. To do so, we propose a cross-attention model that learns relational directions, which are expected displacements between actions and their corresponding objects, to measure their compatibilities. Our framework has several advantages over the state of the art:

– To learn action-object spatial relations, we design a novel cross-attention mechanism, that estimates distributions of relational directions for localizing objects/actions in interactions. Cross-attention is differentiable, which enables its efficient training by backpropagating gradients from interaction scores without requiring bounding-box annotations. The object/action scores computed according to relational directions measure the spatial compatibility between object and action in interaction labels.

– We use the observation that an action-object spatial configuration often depends on the action type. For example, the object location for the action 'sit' is below the action location regardless of the object (e.g., chair, bed), see Figure 1. Thus, we condition our cross-attention predictions on action types in each interaction label, which enables generalization to unseen labels with similar actions.

– Instead of relying on costly bounding-box supervision, we leverage the point-wise localization ability of the visual attention on actions and objects in interaction labels, which can generalize to unseen actions and scale to thousands of labels in the Visual Genome dataset.

2. Related Works

Multi-label Learning addresses recognition of all concepts, such as action, object and attribute labels, in an image [30, 31]. Although it can be addressed by learning a binary classifier for each label [32], this naive approach performs poorly on many labels with insufficient training samples [10, 33]. Thus, the majority of multi-label learning works aims at capturing label dependencies to share their information [34, 11, 12, 13] via label embedding [35, 36], graph neural networks [10, 37, 38, 39, 40], recurrent networks [31, 41] and attention mechanisms [17, 18, 19]. However, they require training samples of every label and cannot recognize unseen labels without training samples.

Zero-Shot Learning aims at recognizing unseen concepts without training samples [42, 43, 44, 45, 46, 47, 48, 49] by leveraging label semantics. Some works [50, 51, 52, 53, 54] further use temporal information to recognize unseen actions in videos. However, most works can only recognize a single unseen label per image. [1, 55, 56, 2] extends zero-shot learning to the multi-label setting. Recently, [3] proposes to share spatial information between labels via attention maps to focus on unseen labels without bounding-box supervision. However, these works target only simple action/object labels, thus do not model action-object spatial relations in interaction labels. Although [9, 57, 58] recognize interaction labels, they do not capture and transfer spatial relation knowledge from seen to unseen interactions.

Spatial Relations in Human-Object Interaction, which are the relative positions between actions and objects in interactions [59, 60, 61, 62, 63], are robust to appearance variations in interactions, hence, have been shown to improve HOI recognition. Thus, recent works focus on detecting humans and objects in actions by relying on bounding-box annotations for each interaction label [6, 64, 65, 66, 67, 68]. [69, 70, 71, 72] convert predicted bounding boxes into binary images to measure their interactiveness scores, while [8, 73, 74] regress object locations from human regions or center of both human and object regions [75]. To reduce annotation costs, [26, 76, 29, 28, 27, 25] propose to detect unseen interaction labels based on visual-spatial information from seen labels. However, these methods require bounding-box annotations to detect interactions. Orthogonal to these works in HOI detection, our method focuses on the *recognition task of unseen interaction labels using solely image-level supervision*.

Weakly-Supervised Localization [77, 78, 79] has recently gained traction due to the high cost of bounding-box annotations needed to train object detectors [80, 81]. [82, 83, 84] discovers that activations of CNNs can be analyzed to infer

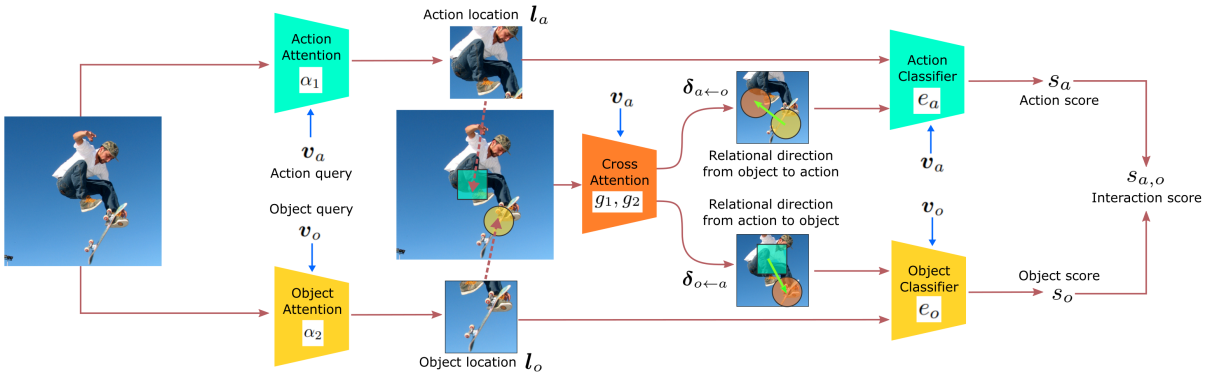

Figure 2: Given an input image I, we first estimates action and object locations, $\boldsymbol{l}_a$, $\boldsymbol{l}_o$ using attention mechanisms. Our proposed cross-attention component then estimates the relational directions from action to object, $\boldsymbol{\delta}_{o \leftarrow a}$, and from object to action, $\boldsymbol{\delta}_{a \leftarrow o}$. Action and object scores, s_a, s_o, are computed based on attention locations and relational directions to capture spatial relations between actions and objects. We compute the interaction score, $s_{a,o}$, as the sum of action and object scores.

locations of objects in images. Recent works [79, 85, 86] focus on capturing the full spatial extents of objects by regularizing predictions to prevent localizing only the most discriminative parts. However, these works cannot localize both actions and objects in interaction labels. Moreover, they require at least image-level supervision for localization, thus cannot generalize to unseen interactions.

3. Compositional Learning of Unseen Interaction Labels using Spatial Relations

In this section, we develop a compositional learning framework for multi-label zero-shot HOI recognition that captures spatial relations. We assume training images are annotated with their ground-truth list of interaction labels without having bounding-box regions of humans or objects.

3.1. Problem Settings

Let $\mathcal{A}$ and $\mathcal{O}$ denote the sets of human action and object labels, respectively. The Cartesian product of the two sets, $\mathcal{A} \times \mathcal{O} \triangleq \{(a, o) |\ a \in \mathcal{A}, o \in \mathcal{O}\}$, corresponds to all possible interaction labels. Throughout the paper, we use the term interaction component to refer to an action or object. In the multi-label zero-shot interaction learning, we have two sets $\mathcal{C}_s, \mathcal{C}_u \subset \mathcal{A} \times \mathcal{O}$, where $\mathcal{C}_s$ corresponds to *seen interaction labels* that have training samples and $\mathcal{C}_u$ denotes *unseen interaction labels* that lack training samples. Let $(I_1, \boldsymbol{Y}_1), \ldots, (I_N, \boldsymbol{Y}_N)$ be N training samples, where I_i denotes the training image i and $\boldsymbol{Y}_i \in \{0, 1\}^{|\mathcal{A}| \times |\mathcal{O}|}$ encodes its ground-truth interactions. An image may contain one or multiple interactions, e.g., the image in Figure 1 contains 'sit + chair' and 'type + laptop'.

The goal of multi-label zero-shot HOI recognition is to classify both seen and unseen interaction labels $(a, o) \in \mathcal{C}_s \cup \mathcal{C}_u$ given only training samples from $\mathcal{C}_s$. Unseen interaction labels correspond to either combinations of seen actions and objects but in a novel way not present in training data or combination of actions and objects at least one of which is unseen. We use word embeddings of actions and objects, $\{\boldsymbol{v}_a^t\}_{a \in \mathcal{A}}, \{\boldsymbol{v}_o^t\}_{o \in \mathcal{O}}$, to handle recognition of unseen interaction labels by leveraging semantic similarities between unseen and seen interaction components.

3.2. Proposed Framework

To address the problem of multi-label zero-shot HOI recognition, we develop a compositional framework in which the interaction score $s_{a,o}$ between an action a and an object o is decomposed as the sum of scores of the action, s_a, and the object, s_o (see Figure 2). As we show, this allows us to transfer the learned knowledge from seen actions and objects to unseen interaction labels.

To compute the action and object scores, it is necessary to localize them in images. Thus, we use *two attention modules* that, in a weakly-supervised setting, learn to select relevant action and object regions to extract attention features. We use the *attention features* to compute action and object *visual scores*. However, given that an image may have multiple actions and objects, combining visual scores cannot capture which actions and objects interact with each other.

Leveraging the spatial dependencies between interaction components, we propose a novel cross-attention mechanism to learn two *relational directions*: one from action to object, which predicts the location of the object based on the action information, and one from object to action, which predicts the location of the action based on the object information. We use these directions to compute *relational features* and subsequently *relational scores* for the actions and objects. Finally, we compute the action/object score as the combination of visual and relational scores, which allow us to more effectively recognize interaction labels. For example, when an action and object exist in an image, hence, have high visual scores, but do not interact, the relational scores would be low/negative, thereby, reducing the interaction score. On the other hand, an action or an object in interaction whose

attention feature is not sufficiently informative (e.g., due to occlusions) will have low visual scores, while the relational features can capture their presence by producing positive relational scores, which increases the interaction score.

3.2.1 Learning Visual Representations of Actions and Objects via Attention Models

To effectively transfer knowledge from seen to unseen interaction labels, we use a compositional learning paradigm. We decompose learning the interaction model into learning action and object models, the combinations of whose outputs allows us to recognize seen and unseen interaction labels. Given the lack of bounding-box interaction annotations and to learn action and object features that encode information from relevant image regions, we use soft-attention [15, 16, 42] to select regions according to a query vector $\boldsymbol{v}$. Let $\{\boldsymbol{f}^r\}_{r=1}^R$ be the region features of an image I, which is divided into R equal-size regions. We compute the attention weights,

$$\alpha(\boldsymbol{f}^r, \boldsymbol{v}) = \frac{\exp(\boldsymbol{v}^T \boldsymbol{W}^\alpha \boldsymbol{f}^r)}{\sum_{r'} \exp(\boldsymbol{v}^T \boldsymbol{W}^\alpha \boldsymbol{f}^{r'})}, \tag{1}$$

where $\boldsymbol{W}^\alpha$ is a learnable matrix that measures the compatibility between the region r of the image and query vector $\boldsymbol{v}$. Thus, α indicates the importance score of each region, normalized by the softmax operation, with respect to $\boldsymbol{v}$.

We use two attention modules, $\alpha_1(\cdot)$ and $\alpha_2(\cdot)$ parametrized by $\boldsymbol{W}_1^\alpha$ and $\boldsymbol{W}_2^\alpha$, to select regions and extract features for actions and objects, respectively,

$$\boldsymbol{h}_a \triangleq \sum_r \alpha_1(\boldsymbol{f}^r, \boldsymbol{v}_a)\boldsymbol{f}^r, \;\; \boldsymbol{h}_o \triangleq \sum_r \alpha_2(\boldsymbol{f}^r, \boldsymbol{v}_o)\boldsymbol{f}^r. \tag{2}$$

Here, $\boldsymbol{h}_a$ and $\boldsymbol{h}_o$ are the attention features for action a and object o, respectively, by aggregating image region features using the attention scores. Query vectors $\boldsymbol{v}_a, \boldsymbol{v}_o$ of actions and objects are used to guide the two attention modules.

Visual Queries. To compute the query vectors, we propose to modify the word embeddings of actions $\{\boldsymbol{v}_a^t\}$ and objects $\{\boldsymbol{v}_o^t\}$ learned from textual corpus (e.g., Wikipedia) into suitable visual queries as

$$\boldsymbol{v}_a = \boldsymbol{v}_a^t + r(\boldsymbol{v}_a^t), \;\; \boldsymbol{v}_o = \boldsymbol{v}_o^t + r(\boldsymbol{v}_o^t), \tag{3}$$

where $r(\cdot)$ is a neural network computing the modifications needed to construct visual query vectors $\boldsymbol{v}_a, \boldsymbol{v}_o$ from their word embedding $\boldsymbol{v}_a^t, \boldsymbol{v}_o^t$. By using the same $r(\cdot)$ for both action and object queries, we effectively share knowledge between interaction components to refine word embeddings into suitable queries instead of learning from scratch.

Naive Approach. Given the attention features for actions and objects, a naive approach to multi-label HOI recognition is to compute the score of interactions as the sum of *visual scores* of actions $e_a(\boldsymbol{h}_a)$ and objects $e_o(\boldsymbol{h}_o)$ from visual attention features,

$$s_{a,o} = e_a(\boldsymbol{h}_a) + e_o(\boldsymbol{h}_o), \tag{4}$$

$$e_a(\boldsymbol{h}) \triangleq \boldsymbol{v}_a^\top \boldsymbol{W}_1^e \boldsymbol{h}, \;\; e_o(\boldsymbol{h}) \triangleq \boldsymbol{v}_o^\top \boldsymbol{W}_2^e \boldsymbol{h}, \tag{5}$$

where $\boldsymbol{W}_1^e, \boldsymbol{W}_2^e$ are embedding matrices to compute the compatibility of action and object attention features with the query vectors $\boldsymbol{v}_a, \boldsymbol{v}_o$. Without incorporating any relations between actions and objects, (4) effectively assumes actions and objects appear independently in an image, thus, cannot distinguish between background objects and objects in interactions (see Figure 1). In the next section, we develop a novel cross-attention mechanism that estimates relational directions and scores to capture spatial dependencies between interaction components while maintaining their compositionality for efficient prediction.

3.2.2 Modeling Spatial Relations between Actions and Objects with Cross Attention

To capture dependencies between interaction components, we use the observation that action and object in an interaction label follow a specific spatial configuration, which often depends on the action type. For example, object location for the action 'sit' must be below the action location regardless of the object type, e.g., 'chair', 'bed' (see Figure 1). Thus, we add to our model a cross-attention component that predicts the location of the object/action based on the location of the action/object using the action query vector. We will use the cross-attention locations to build visual relational features of actions and objects, which we then use in conjunction with (5) to compute interaction scores.

Our first step is estimating action and object locations based on image regions selected by the attention models,

$$\boldsymbol{l}_a \triangleq \sum_r \alpha_a(\boldsymbol{f}^r, \boldsymbol{v}_a)\boldsymbol{l}^r, \;\; \boldsymbol{l}_o \triangleq \sum_r \alpha_o(\boldsymbol{f}^r, \boldsymbol{v}_o)\boldsymbol{l}^r, \tag{6}$$

where $\boldsymbol{l}^r$ is the 2D coordinate of the center of region r and $\boldsymbol{l}_a, \boldsymbol{l}_o$ denote centers of actions and objects, respectively. Notice that we use the sum of locations, weighted by attention scores, as it is fully differentiable for training.

Next, we use the information of action a to predict the location of object o, denoted by $\boldsymbol{l}_{o \leftarrow a}$, and use the information of object o to predict the location of action a, denoted by $\boldsymbol{l}_{a \leftarrow o}$. To do so, we learn relational directions, $\boldsymbol{\delta}_{o \leftarrow a}, \boldsymbol{\delta}_{a \leftarrow o}$, between the two components. More specifically, given attention features and locations of the action $(\boldsymbol{h}_a, \boldsymbol{l}_a)$ and object $(\boldsymbol{h}_o, \boldsymbol{l}_o)$, our cross attention models the displacement from a to o and vice versa using Gaussian distributions,

$$\boldsymbol{\delta}_{o \leftarrow a}, \boldsymbol{\Sigma}_{o \leftarrow a} = g_1\Big(\boldsymbol{h}_a, \boldsymbol{l}_a | \boldsymbol{v}_a\Big), \;\; \boldsymbol{l}_{o \leftarrow a} = \boldsymbol{l}_a + \boldsymbol{\delta}_{o \leftarrow a}, \tag{7}$$

$$\boldsymbol{\delta}_{a \leftarrow o}, \boldsymbol{\Sigma}_{a \leftarrow o} = g_2\Big(\boldsymbol{h}_o, \boldsymbol{l}_o | \boldsymbol{v}_a\Big), \;\; \boldsymbol{l}_{a \leftarrow o} = \boldsymbol{l}_o + \boldsymbol{\delta}_{a \leftarrow o}, \tag{8}$$

where $g_1(\cdot), g_2(\cdot)$ are two neural networks whose outputs are the estimated mean ($\boldsymbol{\delta}$) and covariance ($\boldsymbol{\Sigma}$) of displacements. Notice that the expected location $\boldsymbol{l}_{o\leftarrow a}$ of object o using action a information is the sum of action location $\boldsymbol{l}_a$ and the mean displacement from action to object $\boldsymbol{\delta}_{o\leftarrow a}$ (similarly for object to action). Compared to directly predicting absolute locations of actions and objects, relational directions $\boldsymbol{\delta}_{o\leftarrow a}, \boldsymbol{\delta}_{a\leftarrow o}$ have the advantage of capturing relative visual relationships such as "on top" or "in front of".

Remark 1 *Relational directions depend on action types, encoded in $\boldsymbol{v}_a$, but not on object types, since action types mostly dictate relative locations of actions and objects in interactions. This enables our framework to transfer knowledge of relational directions from seen to unseen interaction labels using similarity of actions regardless of object types.*

Given the expected locations, $\boldsymbol{l}_{o\leftarrow a}, \boldsymbol{l}_{a\leftarrow o}$, we compute relational features of actions and objects from which we compute relational scores. We assume when an interaction labels of a and o occurs, given $\boldsymbol{l}_a$, the relational direction $\boldsymbol{\delta}_{o\leftarrow a}$ would point to the image region containing o, hence, producing a high object score e_o (similarly for action a). On the other hand, the expected locations for unrelated actions and objects would be irrelevant regions that have small or even negative scores. To compute the relational visual features, we weight image region features using the Gaussian probability of their coordinates $\boldsymbol{l}^r$ based on estimated parameters for actions $(\boldsymbol{l}_{a\leftarrow o}, \boldsymbol{\Sigma}_{a\leftarrow o})$ and objects $(\boldsymbol{l}_{o\leftarrow a}, \boldsymbol{\Sigma}_{o\leftarrow a})$ as,

$$\boldsymbol{h}_{o\leftarrow a} \triangleq \sum_r p_g(\boldsymbol{l}^r | \boldsymbol{l}_{o\leftarrow a}, \boldsymbol{\Sigma}_{o\leftarrow a}) \boldsymbol{f}^r, \tag{9}$$

$$\boldsymbol{h}_{a\leftarrow o} \triangleq \sum_r p_g(\boldsymbol{l}^r | \boldsymbol{l}_{a\leftarrow o}, \boldsymbol{\Sigma}_{a\leftarrow o}) \boldsymbol{f}^r, \tag{10}$$

where $p_g(\cdot)$ is the Gaussian density function. Here, $\boldsymbol{h}_{o\leftarrow a}, \boldsymbol{h}_{a\leftarrow o}$ denote the relational features from which we compute the *relational scores* for objects and actions as $e_o(\boldsymbol{h}_{o\leftarrow a})$ and $e_a(\boldsymbol{h}_{a\leftarrow o})$, respectively, using the embedding functions in (5). Notice that the covariance matrices $\boldsymbol{\Sigma}_{o\leftarrow a}, \boldsymbol{\Sigma}_{a\leftarrow o}$ capture the uncertainty of relational directions, where large variance reduces region probabilities leading to small relational scores. Thus, our framework down-weights uncertain predictions. Moreover, inferring the distributions of $\boldsymbol{\delta}_{o\leftarrow a}, \boldsymbol{\delta}_{a\leftarrow o}$ makes our framework differentiable, since the influence of each image region with respect to relational directions changes smoothly, as opposed to predicting a specific image region.

3.2.3 Interaction Label Prediction via Visual Representations and Spatial Relations

To produce the final prediction, we combine the relational scores $e_a(\boldsymbol{h}_{a\leftarrow o}), e_o(\boldsymbol{h}_{o\leftarrow a})$ with the visual scores $e_a(\boldsymbol{h}_a), e_o(\boldsymbol{h}_o)$ to compute the overall scores for actions, denoted by s_a, and objects, denoted by s_o,

$$s_a(\boldsymbol{h}_a, \boldsymbol{h}_{a\leftarrow o}) = e_a(\boldsymbol{h}_a) + w_1 e_a(\boldsymbol{h}_{a\leftarrow o}), \tag{11}$$

$$s_o(\boldsymbol{h}_o, \boldsymbol{h}_{o\leftarrow a}) = e_o(\boldsymbol{h}_o) + w_2 e_o(\boldsymbol{h}_{o\leftarrow a}), \tag{12}$$

where w_1, w_2 are learnable scalars adjusting the relative effect of the two terms. Here, the relational scores modulate the overall scores according to action-object spatial relations. When the relational directions point towards the right regions of the action/object, the relational scores would be high, increasing the overall scores. Otherwise, for an incompatible (action, object) pair, the relational score would be small or negative, suppressing the overall scores.

Finally, we compute the interaction score as the sum of the overall action and object scores,

$$s_{a,o} \triangleq s_a(\boldsymbol{h}_a, \boldsymbol{h}_{a\leftarrow o}) + s_o(\boldsymbol{h}_o, \boldsymbol{h}_{o\leftarrow a}). \tag{13}$$

This allows us to maintain the compositional structure between actions and objects and recombine the learned knowledge to predict the scores of unseen interaction labels.

Loss Function. To train all components of our framework, for each training image, we use the binary cross-entropy loss between interaction scores $s_{a,o}$ and their corresponding ground-truth annotations $y_{a,o}$,

$$\mathcal{L} \triangleq -\sum_{(a,o)\in\mathcal{C}_s} y_{a,o} \log(\sigma(s_{a,o})) + (1 - y_{a,o}) \log(1 - \sigma(s_{a,o})), \tag{14}$$

and minimize the average loss over training images via stochastic gradient descent. Here, $\sigma(\cdot)$ denotes the sigmoid function converting interaction scores to prediction probabilities. We minimize the loss with respect to parameters of the action and object models, $\{\boldsymbol{W}_i^\alpha, \boldsymbol{W}_i^e, g_i, w_i\}_{i=1}^2$, and the visual query model, r.

4. Experiments

We evaluate our proposed framework, which we refer to as Interaction Compass (`ICompass`), for multi-label zero-shot HOI recognition on HICO [7] and Visual Genome [59] datasets. We also analyze the pointwise localization performance [83, 3] by measuring whether predicted locations of actions and objects are within their ground-truth bounding-boxes, on HICO-DET [70]. Unlike weakly-supervised object detection [82, 87, 88], which require training samples for every label, and zero-shot object detection [89, 90, 91], which need bounding-box annotations of seen labels, our setting measures performance of unseen label recognition without bounding-box supervision and training samples. We first discuss the datasets, evaluation metrics, implementation details and baselines. We then present recognition

and localization performances. Finally, we show the effectiveness of the cross-attention for estimating interaction regions and conduct ablation studies to show the necessity of each proposed component.

4.1. Experimental Setup

Datasets. Following [9], we report the zero-shot recognition performance on HICO [7] and Visual Genome [59] datasets, which contain images of various interactions between human and objects. HICO has 38,116 training images and 9,658 testing images carefully collected for 520 interactions from 117 actions and 80 objects. On the other hand, Visual Genome is a visual relation dataset consisting of 520 human actions with 1,422 objects among 21,256 images. This results in 6,643 interactions for training and 532 interactions having at least 10 samples for reliably evaluating the performances.

Similar to [9], we divide the action set into two disjoint sets $A, B \subset \mathcal{A}$ such that $A \cap B = \varnothing$ and similarly divide objects into $1, 2 \subset \mathcal{O}$ where $1 \cap 2 = \varnothing$. Given these sets, we partition interaction labels into 4 sets $A1, B1, A2, B2$, by combining actions and objects from their respective sets, e.g., $A1 \subseteq A \times 1$. We setup two evaluation settings:

(1) Seen interactions: $A1 \cup B2$, Unseen interactions: $B1 \cup A2$,

(2) Seen interactions: $A1$, Unseen interactions: $B1 \cup A2 \cup B2$.

Notice that Setting 1 tests the ability to recombine knowledge from seen interactions, since all actions and objects are observed, while Setting 2 requires to extrapolate to unseen actions and objects. Due to a large number of interactions, both datasets contain missing annotations in images, which are treated as negative labels, similar to [7, 9].

Evaluation Metrics. Following other works on multi-label learning [7, 9, 3], we measure the mean Average Precision (mAP) capturing how well a model retrieves relevant samples for each interaction. We also report the ranking measurement, F1 score, which is the harmonic mean between the precision and recall of top-10 predictions in each image. Notice that mAP compares predictions across different images while F1 distinguishes between interactions within the same image, hence, these measurements offer complementary performance information.

Baselines. We compare with `GCNCL` [9], which exploits external knowledge graphs based on WordNet between actions and objects to construct unseen interaction classifiers. Following [9], we further compare with methods using only image-level labels. Thus, we adapt multi-label zero-shot learning works [92, 23, 3] to recognize interaction labels by predicting action and object scores and adding them into corresponding interaction scores. To be specific, we employ `DEVISE` [92], which learns linear embedding spaces, and `Fast0Tag` [23], which constructs a nonlinear embedding function to measure the compatibility between image features and word embeddings of actions or objects. We also use `LESA` [3], as the state-of-the-art multi-label zero-shot learning model, which learns to share attention among related action/object labels.

To show the importance of visual query refinement and spatial relations, we consider a `Dual Attention` baseline, consisting of two independent soft-attention modules for actions and objects using word embeddings as queries without learning spatial relations. As an attempt to capture action-object dependencies, we construct a `Combined Attention` baseline, which extends `Dual Attention` by predicting dependency scores from the concatenation of visual features $\boldsymbol{h}_a, \boldsymbol{h}_o$ and locations $\boldsymbol{l}_a, \boldsymbol{l}_o$, parametrized by a neural network. Since `Combined Attention` heavily relies on the correctness of both action and object attention predictions in each interaction to compute dependency scores, it will not be robust against incorrect localization in either components. Our method only depends on either action or object locations to infer spatial relations and captures prediction uncertainties, thus can also correct localization errors (see the supplementary materials).

Implementation Details. To extract region features for attention mechanism, we use the feature map from the last convolutional layer of a pretrained ResNet-152 whose size is $W \times H \times 2048$ and treat it as a set of features from $W \times H$ regions. We pad input images, such that they have equal widths and heights, and reshape them into 544×544 size and 17×17 image regions, which achieves good trade-off between performances and memory consumptions. For each region, we assign a unique 2D coordinate $\boldsymbol{l}^r$ in the range of $[1, 17] \times [1, 17]$. We parametrize $g_1(\cdot), g_2(\cdot)$ in cross attention as two neural networks with one hidden layer of size 300. We normalize relational directions within image ranges and predict positive diagonal matrices for the covariance matrices. Similarly, for the visual query, $r(\cdot)$ is model as a neural network with one hidden layer of size 60. We extract the semantic vectors $\{\boldsymbol{v}_a^t\}_{a \in \mathcal{A}}, \{\boldsymbol{v}_o^t\}_{o \in \mathcal{O}}$ using the GloVe model [93] trained on Wikipedia articles. We implement all methods in PyTorch and optimize using RMSprop [94] with its default setting, learning rate 0.001 and batch size of 32 on 10 epochs on all datasets.

4.2. Experimental Results

Multi-Label Zero-Shot HOI Recognition. We report performances on only unseen interaction labels (Unseen), and on both seen and unseen interaction labels (All) corresponding to, respectively, zero-shot and generalized zero-shot settings. Table 1 shows F1 scores at top-10 predictions and mAP scores for all methods in both Setting 1 ($A1 \cup B2$ setting) and Setting 2 ($A1$ setting). From the results, we make the following conclusions:

Method	Seen Interactions	HICO						Visual Genome					
		Unseen			All			Unseen			All		
		R@10	F1@10	mAP	R@10	F1@10	mAP	R@10	F1@10	mAP	R@10	F1@10	mAP
GCNCL* [9]	*A1* ∪ *B2*	-	-	17.0	-	-	21.4	-	-	5.4	-	-	6.6
	A1	-	-	7.5	-	-	11.9	-	-	2.4	-	-	4.1
ICompass (Ours)*	*A1* ∪ *B2*	78.1	12.7	**19.8**	76.1	25.2	**25.9**	52.0	5.7	**6.9**	47.5	10.2	**7.8**
	A1	38.2	9.8	**8.9**	43.1	14.2	**14.7**	25.8	4.0	**3.4**	29.7	6.4	**4.7**
DEVISE [92]	*A1* ∪ *B2*	54.4	8.8	10.7	59.0	19.5	16.9	35.1	3.9	3.0	31.5	6.8	3.7
	A1	15.8	4.1	3.7	22.6	7.5	8.1	6.3	1.0	1.4	14.0	3.0	2.1
Fast0Tag [23]	*A1* ∪ *B2*	76.8	12.5	19.9	75.8	25.1	26.2	49.1	5.4	7.1	43.9	9.5	8.0
	A1	41.5	10.7	8.7	**48.1**	**15.9**	14.8	25.1	3.8	3.9	**33.3**	**7.2**	5.1
LESA [3]	*A1* ∪ *B2*	71.1	11.5	21.8	75.0	24.8	28.3	41.6	4.6	6.9	43.2	9.3	8.3
	A1	29.8	7.7	9.8	40.1	13.3	16.2	7.7	1.2	2.9	14.7	3.2	4.4
Dual Attention	*A1* ∪ *B2*	71.2	11.6	19.1	73.9	24.5	25.8	51.2	5.7	6.1	47.8	10.3	7.5
	A1	32.5	8.4	9.1	41.4	13.7	15.1	16.4	2.5	3.0	28.1	6.1	4.1
Combined Attention	*A1* ∪ *B2*	71.1	11.5	14.3	72.0	23.8	22.1	41.9	4.6	4.9	40.8	8.8	5.9
	A1	27.7	7.1	7.7	38.2	12.6	13.0	18.6	2.9	2.6	28.3	6.1	3.6
ICompass (Ours)	*A1* ∪ *B2*	**82.7**	**13.4**	**24.4**	**81.6**	**27.0**	**30.4**	**57.0**	**6.3**	**7.8**	**52.1**	**11.2**	**8.8**
	A1	**42.0**	**10.8**	**10.8**	46.6	15.4	**17.2**	**26.8**	**4.1**	**4.2**	31.4	6.8	**5.4**

Table 1: Multi-label zero-shot HOI recognition performance on HICO/Visual Genome. * indicates 224×224 image resolution input.

Method	Seen Interactions	Action					Object					Action & Object				
		A1	*A2*	*B1*	*B2*	All	*A1*	*A2*	*B1*	*B2*	All	*A1*	*A2*	*B1*	*B2*	All
DEVISE [92] + CAM [84]	*A1* ∪ *B2*	9.7	6.0	9.6	10.8	9.0	11.6	11.6	12.6	18.5	13.6	4.6	4.2	4.6	7.1	5.2
	A1	5.4	3.9	3.4	2.3	3.8	8.3	2.9	5.0	**2.1**	4.6	3.1	1.5	1.6	0.9	1.8
LESA [3]	*A1* ∪ *B2*	23.6	19.9	18.8	21.8	21.0	24.8	26.3	19.4	25.6	24.0	13.3	15.2	9.8	13.5	13.0
	A1	23.6	4.0	18.8	1.3	11.9	25.0	4.0	18.8	1.6	12.4	12.9	2.6	10.2	1.0	6.7
Dual Attention	*A1* ∪ *B2*	21.0	16.9	20.6	22.3	20.2	21.6	20.6	20.0	24.0	21.5	10.9	12.5	11.6	14.1	12.3
	A1	21.6	**6.8**	15.4	1.4	11.3	23.4	3.4	13.6	0.5	10.2	11.7	2.6	7.7	0.4	5.6
Combined Attention	*A1* ∪ *B2*	17.9	6.7	20.3	19.8	16.2	13.2	9.3	12.0	12.9	11.9	5.9	3.7	6.7	6.7	5.7
	A1	17.2	6.5	11.8	**2.6**	9.5	12.4	2.5	7.9	0.9	5.9	5.5	1.9	4.0	0.8	3.0
ICompass (Ours)	*A1* ∪ *B2*	**28.6**	**23.5**	**28.4**	**28.1**	**27.2**	**32.4**	**29.7**	**25.9**	**33.6**	**30.4**	**17.5**	**18.4**	**15.9**	**19.6**	**17.9**
	A1	**28.0**	6.1	**19.8**	1.1	**13.7**	**34.4**	**5.8**	**20.0**	1.1	**15.3**	**15.9**	**4.0**	**11.3**	**1.0**	**8.0**

Table 2: Zero-shot HOI localization (mAP) performance on images having the target interaction labels in the HICO-DET dataset.

– In GCNCL setup with input size of 224×224[2], we significantly outperform GCNCL on both datasets, for both settings 1 and 2 and for both Unseen and All interaction predictions. Thus, without decoupling visual features of actions and objects, GCNCL is unable to capture and transfer information from seen to unseen interaction labels even when relying on external knowledge. In the higher image resolution setup of 544×544, we further boost our performances by 3.6% and 1.9% on HICO in *A1* ∪ *B2* and *A1* settings, respectively, for unseen mAP as our method can attend to finer image regions. We use this resolution in the remaining experiments, as it also benefits other baselines, for fair comparison.

– On unseen interactions, our method surpasses the state of the art on HICO not only by 2.6% and 1.0% mAP scores, but also by 5.9% and 0.5% recall scores on *A1* ∪ *B2* and *A1* settings, respectively, showing that most of our confident predictions are accurate. To evaluate on Visual Genome, due to its missing and noisy labels, we use recall, which does not penalize predictions of unannotated interactions compared to mAP. We achieve at least 5.8% (*A1* ∪ *B2* setting) and 1.7% (*A1* setting) recall improvements and the best F1 and mAP performances.

– On all interactions, we improve mAP scores by 2.1% (1.0%) on HICO for *A1* ∪ *B2* (*A1*) setting and achieve comparable performances to the state of the art on Visual Genome. Although Fast0Tag achieves high ranking performances on all interactions, its low unseen performances indicates the baseline mostly overfits to seen interactions without generalizing to unseen interactions.

Zero-Shot HOI Localization. To further analyze the performance, we propose to measure the localization performances for each interaction label on only images having the target label following [95], please prefer to the supplementary materials for evaluation on all images. We consider top-10 predictions in each image as positive predictions and, using mAP scores [83, 3], we measure whether the ground-truth labels are within top predictions and their predicted locations[3] are within the ground-truth bounding boxes on HICO-DET dataset [70]. We evaluate action localization, object localization, and action-object localization where a model need to correctly localize both components. Here, we use human bounding boxes as a proxy for locations of actions which are performed by humans. Table 2 shows the results, which support the following conclusions:

– Overall, most methods localize objects better than actions, while action-object localization is the most difficult task. In *A1* ∪ *B2* setting, our method achieves the best localization performances on both seen and unseen interactions compared to the state of the art, resulting in 6.2%, 6.4% and 4.9% all improvement across action, object and action-object localization, respectively.

– In the *A1* setting, most methods perform better at local-

[2] As we are unable to obtain the code from authors, we use their reported mAP scores to ensure reporting their best performances.

[3] We use Class Activation Map [84] to locate actions/objects for DEVISE which lacks localization ability.

Figure 3: Visualization of object/action attention maps and object location distributions estimated by cross attention for unseen interaction labels.

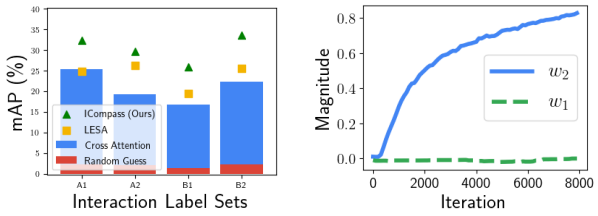

Figure 4: Left: Object localization score of cross-attention. Right: Changes in learnable weights, w_1, w_2, for relational scores across training iterations. Experiments are performed in *A1* ∪ *B2* setting on HICO-DET.

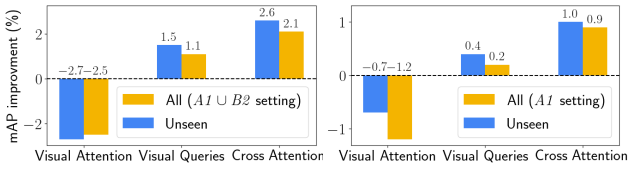

Figure 5: Multi-label zero-shot HOI recognition improvements over LESA on HICO when adding each component of our method.

izing interaction labels of seen than unseen objects, which shows that object features are highly discriminative for recognizing interactions. Overall, our methods significantly improve mAP scores by 1.8%, 2.9%, 1.3% on action, object and action-object pointwise localization, respectively.

– `Combined Attention` has low performance since it naively combines action and object predictions without accounting for localization errors, thus propagates the errors and degrades the performance. The modest performance of `Dual Attention` shows the importance of capturing action/object spatial relations in addition to visual attention.

Effectiveness of Cross Attention. To show that objects can be located from actions' information, we compare pointwise object localization performances (mAP) of cross-attention, $\boldsymbol{l}_{o \leftarrow a}$, shared attention in `LESA` and `ICompass`, shown in Figure 4 (left). Compared to uniformly select regions (`Random Guess`), cross attention is significantly better at object localization, thus verifies the effectiveness of using action information to estimate relational directions.

We also visualize the learnable weights for relational scores (w_1, w_2) across training iterations in Figure 4 (right). During training, our model gradually uses relational directions from actions ($w_2 = 0.8$) and suppresses directions from object ($w_1 = 0.0$). As objects do not significantly change in interactions, e.g., 'cup' remains visually unchanged under 'pouring into' or 'drinking from', their features are unreliable for predicting spatial relations.

Ablation Studies. We conduct ablation study of the recognition performance (mAP) on the HICO dataset to measure the improvements when adding each component of our framework compared to `LESA`. As shown in Figure 5, we observe improvements using our visual queries (see the supplementary materials for query visualization), which shares knowledge between object and action queries, while `LESA` learns them independently. Cross-attention significantly boosts the performance by 1% in *A1* ∪ *B2* setting and doubles the improvements in *A1* setting on both unseen and all interactions. This shows our method succeeds in transferring knowledge from seen to unseen interactions.

Qualitative results. Figure 3 visualizes attention maps of actions/objects and distributions of object locations $(\boldsymbol{l}_{o \leftarrow a}, \boldsymbol{\Sigma}_{o \leftarrow a})$ from cross attention on HICO-DET dataset. Our method can focus on relevant regions of different actions, such as 'ride' and 'pet', to recognize multiple unseen interaction labels in the *A1* ∪ *B2* setting. Moreover, cross-attention successfully attends to objects corresponding to each action. In *A1* setting, we generalize to the unseen action 'hold' to produce a low relational score for 'ball' whose location is incompatible with our cross-attention prediction.

5. Conclusions

We proposed a compositional multi-label zero-shot interaction learning framework that decouples and recombines action and object knowledge to recognize seen/unseen interactions. We introduced a novel cross-attention model that captures spatial relations between actions and objects to determine their compatibility without bounding-box annotations. Extensive experiments on HICO and Visual Genome datasets demonstrated our ability to recognize unseen interactions, provide estimation of interaction locations and generalize to interaction labels with unseen actions.

Acknowledgements

This work is supported by NSF (IIS-2115110), DARPA Young Faculty Award (D18AP00050), ONR (N000141812132) and ARO (W911NF1810300, W911NF2110276).

References

[1] Y. Zhang, B. Gong, and M. Shah, "Fast zero-shot image tagging," *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 2

[2] C. W. Lee, W. Fang, C. K. Yeh, and Y. C. F. Wang, "Multi-label zero-shot learning with structured knowledge graphs," *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1, 2

[3] D. Huynh and E. Elhamifar, "A shared multi-attention framework for multi-label zero-shot learning," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 5, 6, 7

[4] E. Elhamifar and D. Huynh, "Self-supervised multi-task procedure learning from instructional videos," *European Conference on Computer Vision*, 2020. 1

[5] E. Elhamifar and Z. Naing, "Unsupervised procedure learning via joint dynamic summarization," *International Conference on Computer Vision*, 2019. 1

[6] S. Gupta and J. Malik, "Visual semantic role labeling," *ArXiv*, 2015. 1, 2

[7] Y. W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "Hico: A benchmark for recognizing human-object interactions in images," *IEEE International Conference on Computer Vision*, 2015. 1, 5, 6

[8] G. Gkioxari, R. B. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2

[9] K. Kato, Y. Li, and A. Gupta, "Compositional learning for human object interaction," *European Conference on Computer Vision*, 2018. 1, 2, 6, 7

[10] D. Huynh and E. Elhamifar, "Interactive multi-label CNN learning with partial labels," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2

[11] J. Ma and Y. Liu, "Latent topic-aware multi-label classification," *European Conference on Computer Vision*, 2020. 1, 2

[12] J. Li, C. Zhang, P. Zhu, B. Wu, L. Chen, and Q. Hu, "Spl-mll: Selecting predictable landmarks for multi-label learning," *European Conference on Computer Vision*, 2020. 1, 2

[13] J. Ye, J. He, X. Peng, W. Wu, and Y. Qiao, "Attention-driven dynamic graph convolutional network for multi-label image recognition," *European Conference on Computer Vision*, 2020. 1, 2

[14] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," *Neural Information Processing Systems*, 2015. 1

[15] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," 2015. 1, 4

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Neural Information Processing Systems*, 2017. 1, 4

[17] Z. Wang, T. Chen, G. Li, G. Li, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," *IEEE International Conference on Computer Vision*, 2017. 1, 2

[18] S. F. Chen, Y. C. Chen, C. K. Yeh, and Y. C. F. Wang, "Order-free rnn with visual attention for multi-label classification," *AAAI Conference on Artificial Intelligence*, 2018. 1, 2

[19] T. Chen, Z. Wang, G. Li, and L. Lin, "Recurrent attentional reinforcement learning for multi-label image recognition," *AAAI Conference on Artificial Intelligence*, 2018. 1, 2

[20] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," *Neural Information Processing Systems*, 2017. 1

[21] H. Fang, J. Cao, Y.-W. Tai, and C. Lu, "Pairwise body-part attention for recognizing human-object interactions," *European Conference on Computer Vision*, 2018. 1

[22] T. Xiao, Q. Fan, D. Gutfreund, M. Monfort, A. Oliva, and B. Zhou, "Reasoning about human-object interactions through dual attention networks," *IEEE International Conference on Computer Vision*, 2019. 1

[23] Z. Zhang and V. Saligrama, "Zero-shot learning via joint latent similarity embedding," *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2, 6, 7

[24] T. Mensink, E. Gavves, and C. G. Snoek, "Costa: Co-occurrence statistics for zero-shot classification," *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2

[25] L. Shen, S. Yeung, J. Hoffman, G. Mori, and L. Fei-Fei, "Scaling human-object interaction recognition through zero-shot learning," *IEEE Winter Conference on Applications of Computer Vision*, 2018. 2

[26] Z. Hou, X. Peng, Y. Qiao, and D. Tao, "Visual compositional learning for human-object interaction detection," *European Conference on Computer Vision*, 2020. 2

[27] A. Bansal, S. S. Rambhatla, A. Shrivastava, and R. Chellappa, "Detecting human-object interactions via functional generalization," *AAAI Conference on Artificial Intelligence*, 2020. 2

[28] S. Wang, K. H. Yap, J. Yuan, and Y. P. Tan, "Discovering human interactions with novel objects via zero-shot learning," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[29] Y. Liu, J. Yuan, and C. Chen, "Consnet: Learning consistency graph for zero-shot human-object interaction detection," *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 2

[30] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Intenational Journal Data Warehousing and Mining*, vol. 3, 2007. 2

[31] V. O. Yazici, A. Gonzalez-Garcia, A. Ramisa, B. Twardowski, and J. v. d. Weijer, "Orderless recurrent models for multi-label classification," *European Conference on Computer Vision*, 2020. 2

[32] I. Misra, C. L. Zitnick, M. Mitchell, and R. Girshick, "Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels," *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[33] T. Wu, Q. Huang, Z. Liu, Y. Wang, and D. Lin, "Distribution-balanced loss for multi-label classification in long-tailed datasets," *European Conference on Computer Vision*, 2020. 2

[34] L. Jing, L. Yang, and J. Y. M. K. Ng, "Semi-supervised low-rank mapping learning for multi-label classification," *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[35] J. Weston, S. Bengio, and N. Usunier, "Wsabie: Scaling up to large vocabulary image annotation," *IJCAI*, 2011. 2

[36] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, "Deep convolutional ranking for multilabel image annotation," *International Conference on Learning Representations*, 2013. 2

[37] S. Behpour, W. Xing, and B. D. Ziebart, "Arc: Adversarial robust cuts for semi-supervised and multi-label classification," *AAAI Conference on Artificial Intelligence*, 2018. 2

[38] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam, "Large-scale object classification using label relation graphs," *European Conference on Computer Vision*, 2014. 2

[39] Z. M. Chen, X. S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. abs/1904.03582, 2019. 2

[40] T. Chen, M. Xu, X. Hui, H. Wu, and L. Lin, "Learning semantic-specific graph representation for multi-label image recognition," *IEEE International Conference on Computer Vision*, 2019. 2

[41] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[42] D. Huynh and E. Elhamifar, "Fine-grained generalized zero-shot learning via dense attribute-based attention," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2, 4

[43] ——, "Compositional zero-shot learning via fine-grained dense feature composition," *Neural Information Processing Systems*, 2020. 2

[44] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning — the good, the bad and the ugly," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[45] E. Schönfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, "Generalized zero- and few-shot learning via aligned variational autoencoders," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[46] R. Felix, B. G. V. Kumar, I. D. Reid, and G. Carneiro, "Multi-modal cycle-consistent generalized zero-shot learning," *European Conference on Computer Vision*, 2018. 2

[47] H. Jiang, R. Wang, S. Shan, and X. Chen, "Transferable contrastive network for generalized zero-shot learning," *IEEE International Conference on Computer Vision*, 2019. 2

[48] Y. Atzmon and G. Chechik, "Adaptive confidence smoothing for generalized zero-shot learning," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[49] Y. Gong, S. Karanam, Z. Wu, K. Peng, J. Ernst, and P. Doerschuk, "Learning compositional visual concepts with mutual consistency," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[50] D. Mandal, S. Narayan, S. Dwivedi, V. Gupta, S. Ahmed, F. Khan, and L. Shao, "Out-of-distribution detection for generalized zero-shot action recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[51] P. Mettes and C. G. M. Snoek, "Spatial-aware object embeddings for zero-shot localization and classification of actions," *IEEE International Conference on Computer Vision*, 2017. 2

[52] B. Brattoli, J. Tighe, F. Zhdanov, P. Perona, and K. Chalupka, "Rethinking zero-shot video classification: End-to-end training for realistic applications," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[53] A. Mishra, V. Verma, M. K. Reddy, A. Subramaniam, P. Rai, and A. Mittal, "A generative approach to zero-shot and few-shot action recognition," *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018. 2

[54] J. Gao, T. Zhang, and C. Xu, "Learning to model relationships for zero-shot video classification." *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2

[55] M. Ye and Y. Guo, "Multi-label zero-shot learning with transfer-aware label embedding projection," *CoRR*, vol. abs/1808.02474, 2018. 2

[56] Z. Lu, J. Zeng, S. Shan, and X. Chen, "Zero-shot facial expression recognition with multi-label label propagation," *Asian Conference on Computer Vision*, vol. abs/1512.06963, 2018. 2

[57] A. Sarullo and T. Mu, "Zero-shot human-object interaction recognition via affordance graphs," *ArXiv*, 2020. 2

[58] Q. Wang and K. Chen, "Multi-label zero-shot human action recognition via joint latent ranking embedding," *Neural networks*, 2020. 2

[59] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. 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*, 2016. 2, 5, 6

[60] C. Lu, R. Krishna, M. S. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," *European Conference on Computer Vision*, 2016. 2

[61] R. Yu, A. Li, V. Morariu, and L. Davis, "Visual relationship detection with internal and external linguistic knowledge distillation," *IEEE International Conference on Computer Vision*, 2017. 2

[62] B. A. Plummer, A. Mallya, C. M. Cervantes, J. Hockenmaier, and S. Lazebnik, "Phrase localization and visual relationship detection with comprehensive image-language cues," *IEEE International Conference on Computer Vision*, 2017. 2

[63] X. Liang, L. Lee, and E. Xing, "Deep variation-structured reinforcement learning for visual relationship and attribute detection," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[64] W.-s. Z. Hai Wang and L. Yingbiao, "Contextual heterogeneous graph network for human-object interaction detection," *European Conference on Computer Vision*, 2018. 2

[65] T. Wang, R. Anwer, M. H. Khan, F. Khan, Y. Pang, L. Shao, and J. Laaksonen, "Deep contextual attention for human-object interaction detection," *IEEE International Conference on Computer Vision*, 2019. 2

[66] T. Gupta, A. G. Schwing, and D. Hoiem, "No-frills human-object interaction detection: Factorization, layout encodings, and training techniques," *IEEE International Conference on Computer Vision*, 2019. 2

[67] B. Zhuang, L. Liu, C. Shen, and I. Reid, "Towards context-aware interaction recognition for visual relationship detection," *IEEE International Conference on Computer Vision*, 2017. 2

[68] X. Zhong, C. Ding, X. Qu, and D. Tao, "Polysemy deciphering network for robust human-object interaction detection," *European Conference on Computer Vision*, 2020. 2

[69] C. Gao, Y. Zou, and J. B. Huang, "ican: Instance-centric attention network for human-object interaction detection," *IEEE British Machine Vision Conference*, 2018. 2

[70] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," *IEEE Winter Conference on Applications of Computer Vision*, 2018. 2, 5, 7

[71] Y. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H. Fang, Y. Wang, and C. Lu, "Transferable interactiveness knowledge for human-object interaction detection," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[72] C. Gao, J. Xu, Y. Zou, and J. Huang, "Drg: Dual relation graph for human-object interaction detection," *European Conference on Computer Vision*, 2020. 2

[73] Z. Yang, D. Mahajan, D. Ghadiyaram, R. Nevatia, and V. Ramanathan, "Activity driven weakly supervised object detection," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[74] P. Zhou and M. Chi, "Relation parsing neural network for human-object interaction detection," *IEEE International Conference on Computer Vision*, 2019. 2

[75] T. Wang, T. Yang, M. Danelljan, F. Khan, X. Zhang, and J. Sun, "Learning human-object interaction detection using interaction points," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[76] J. Peyre, J. Sivic, I. Laptev, and C. Schmid, "Detecting unseen visual relations using analogies," *IEEE International Conference on Computer Vision*, 2019. 2

[77] T. Durand, T. Mordan, N. Thome, and M. Cord, "Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[78] J. Choe, S. J. Oh, S. Lee, S. Chun, Z. Akata, and H. Shim, "Evaluating weakly supervised object localization methods right," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[79] X. Li, M. Kan, S. Shan, and X. Chen, "Weakly supervised object detection with segmentation collaboration," *IEEE International Conference on Computer Vision*, 2019. 2, 3

[80] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015. 2

[81] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6517–6525, 2017. 2

[82] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2, 5

[83] I. L. M. Oquab, L. Bottou and J. Sivic, "Is object localization for free? – weakly-supervised learning with convolutional neural networks," *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2, 5, 7

[84] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2, 7

[85] K. K. Singh and Y. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," *IEEE International Conference on Computer Vision*, 2017. 3

[86] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. Huang, "Adversarial complementary learning for weakly supervised object localization," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[87] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 5

[88] Z. Ren, Z. Yu, X. Yang, M. Liu, Y. Lee, A. G. Schwing, and J. Kautz, "Instance-aware, context-focused, and memory-efficient weakly supervised object detection," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 5

[89] A. Bansal, K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran, "Zero-shot object detection," *European Conference on Computer Vision*, 2018. 5

[90] S. Rahman, S. Khan, and N. Barnes, "Transductive learning for zero-shot object detection," *IEEE International Conference on Computer Vision*, 2019. 5

[91] P. Zhu, H. Wang, and V. Saligrama, "Don't even look once: Synthesizing features for zero-shot detection," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 5

[92] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," *Neural Information Processing Systems*, 2013. 6, 7

[93] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," *Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 6

[94] T. Tijmen and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning 4.2*, 2012. 6

[95] T. Deselaers, B. Alexe, and V. Ferrari, "Localizing objects while learning their appearance," *European Conference on Computer Vision*, 2010. 7