A Simple Baseline for Weakly-Supervised Human-centric Relation Detection

Raghav Goyal[12]
rgoyal14@cs.ubc.ca
Leonid Sigal[123]
lsigal@cs.ubc.ca

[1] University of British Columbia
Vancouver, BC, Canada
[2] Vector Institute for AI
[3] CIFAR AI Chair

Abstract

In this paper we address the problem of weakly-supervised Visual Relation Detection (VRD) and human-centric Scene Graph generation. Unlike prior works, we assume weaker, yet more natural, supervisory signals. Specifically, we only assume a pre-trained person detector, a generic region proposal mechanism and a set of image-level object and relation labels per frame. Given this data we formulate a very simple architecture with multi-task weak-supervision at object level (for individual proposed regions) and relation level (for each person-object region pair). We show that despite simplicity, our approach achieves state-of-the-art results as compared to other weakly- and strongly-supervised VRD models that are significantly more complex. In ablations, we also show that proposed multi-task learning improves relation predictions. Our goal in this paper is to propose a strong, yet simple, baseline which will spur further developments in the VRD task.

1 Introduction

Object detection has improved significantly with the introduction of neural architectures designed specifically for detection tasks, *e.g.*, Faster RCNN [35], Yolo [34] and, most recently, DETR [62]. However, object detection in itself is inherently limited and is unable to produce detailed semantic representations that are necessary for higher level visual cognition (*e.g.*, visual question answering, human computer interaction, *etc.*). To this end, recent focus has evolved to approaches capable of more holistic and contextual scene understanding. This includes *Visual Relation Detection* (VRD) which aims to detect objects and predict their relationships in an image in the form of <subj,pred,obj> triplets, *e.g.*, person (subj) laying (pred) on a bed (obj), and scene graph generation – collections of <subj,pred,obj> triplets forming a graph-based representation of the scene with nodes corresponding to grounded object instances and directed edges to predicate relations among those objects.

However, to a large extent, most of existing architectures focus on fully supervised scenarios where visual relations, or scene graphs, need to be completely annotated in large scale training datasets (comprising bounding boxes for object instances, object class labels per instance and labeled relations among those instances). Attaining such datasets at scale is undesirable, practically difficult and financially costly. Similar arguments motivated weakly-supervised object detection [1, 3, 42, 43] and hold, even more strongly, for VRD and scene

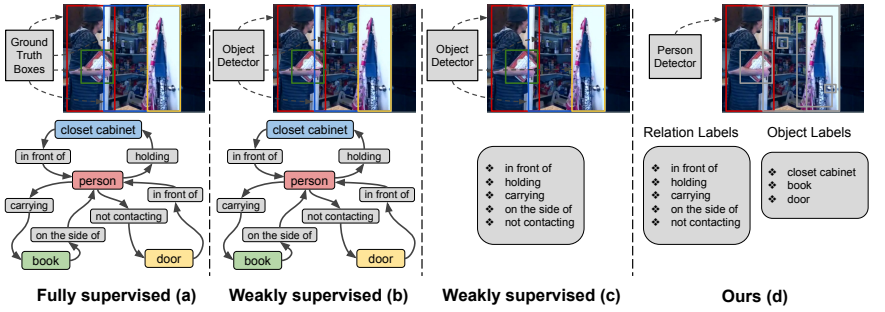

Figure 1: **Forms of HOI task**: Depicted are both supervised and weakly-supervised variants that differ in scale and types of supervision assumed. Fully supervised approaches **(a)** use annotated object boxes and full scene graphs as supervision. Weakly supervised approaches use bounding boxes (and class labels) from a pretrained object detector along with either a full scene graph [29, 57] **(b)** or image-level relation labels [2] **(c)**. Our approach **(d)** uses only a pretrained person detector and (unaligned) sets of image-level relation and object labels.

graph generation, where data is both more complex and less intuitive to annotate. Nevertheless, research on weakly-supervised variants of those tasks has been limited. Few approaches that exist, either assume existence of pre-trained detectors [2] (trained on an external dataset) and/or structured relational labels at the image-level [29, 57] (see Figure 1 (b) and (c)). We argue that neither of those assumptions are ideal in practice. Pre-training detectors themselves requires significant supervision with granular bounding box labels and a class label alignment between external labeled datasets and target data. On the other hand, using structured weak relation labels (or un-localized relation triplets) can lead to combinatorial explosion in annotation effort [2].

In this paper we specifically explore human-object interactions (HOI)[1] with a weaker form of supervision. We only assume image-level object and relations (but unlike [29, 57] no correspondence among them) and a person detector (as opposed to more general object detection [2]). We argue that this setting is significantly simpler from the user and data annotation perspective. Further, we show that simple multi-task weakly-supervised architecture, which simultaneously learns to classify proposal boxes into object classes and pairs of proposals into their predicates, performs well in practice. Despite simplicity, it achieves better performance than prior weakly-supervised variants that rely on more supervision and sophisticated models.

We experiment with benchmark HOI datasets from two different domains: image-based HICO-DET [5] and video-based Action Genome [15]. We provide performance using our weaker supervision and argue that it sets a strong baseline for any future work, and especially in the case of HICO-DET [5], the performance surpasses the state-of-the-art results [2].

Contributions: Our contributions are two fold. First, formulation of what we believe to be a more natural weakly-supervised Visual Relation Detection (VRD) and scene-graph generation task. Second, establishing a simple but surprisingly strong baseline on this task that leverages multiple-instance learning, implemented using identical object- and relation-level losses. We illustrate that this multi-task weakly-supervised setup is beneficial and, overall, results in performance that sets a new weakly-supervised state-of-the-art (despite relying on less supervision and much simplified architectural design). We further show that our results can rival state-of-the-art fully-supervised scene graph generation approaches [20, 28, 52, 53, 56].

[1] HOI is a type of VRD / scene-graph task, that specifically focuses on humans interacting with objects.

2 Related Work

2.1 Visual Relationship Detection

Visual Relationship Detection (VRD) has been extensively studied in the context of general relationships between pairs of objects, ranging from semantic to spatial and comparative correspondences [6, 19, 21, 28, 31, 58, 63]. Related is also the task of scene graph generation, where the focus is typically on a set of densely interconnected and grounded VRD relations, *e.g.*, [20, 52, 53, 55, 56, 59]. A particular instance of VRD where only person is considered as the subject is known as human-object interactions (HOI) [5, 9, 12, 13, 18, 29, 33, 61], which is particularly interesting since most of the rich verbs (*e.g.*, `talk`, `drink`, `throw`) are applicable to person as the subject [12]. However, gathering full annotations for HOI is prohibitively time consuming and expensive. Therefore, in this work we tackle HOI task using weak supervision and follow compositional approach where object and relation detection streams are modelled separately but built on shared feature representations.

2.2 Weakly-supervised object detection

Weakly-supervised learning was initially studied in the context of object detection (WSOD) where the problem was framed as a Multiple Instance Learning (MIL) task [3, 42, 43]. The task involves using image-level supervision to ground an external bag of object proposals [44] to actual objects in an image. Bilen and Vedaldi [3] introduced a framework (named WSDDN) which uses two streams to classify proposals into one of the target classes, while ensuring not many proposals get assigned the same class. This was followed by several improvements in the form of iteratively refining classifiers based on spatial dependence between object parts and locations [42, 43], to training a strong detector based on pseudo-labels from its weak counterpart [1, 46]. In this work, we adopt the simple and intuitive WSDDN framework [3], while noting that our approach is not constrained to this choice.

2.3 Weakly-supervised relation detection

Weakly-supervised relation detection (WSRD) is an abstraction on top of detected objects, where the task is to identify and classify relations between pairs of objects using only image-level relation labels. The task has recently gained attention with only a couple of published works, owing to the difficulty of the problem. Prest et al. [32] used part-based detectors to detect objects relevant to desired action or relation label. Recently, PPR-FCN [57] proposed weakly-supervised object and relation detection modules using WSDDN [3] framework, and assumes image-level triplet annotations as supervision. Specifically, for every ground-truth `<s,r,o>` triplet, they select all the candidate regions for subject `s` and object `o`, and apply weak WSDDN [3] loss over all the candidates for the triplet. Peyre et al. [29] also uses such ground-truth triplets but uses a discriminative clustering scheme for their weakly-supervised module. However, a common theme among the above works is using image-level triplet annotations, which is expensive and suffers from combinatorial explosion when annotating all possible triplets exhaustively [2]. More recently, Baldassarre et al. [2] proposed to use only image-level relation labels, and showed that it leads to a viable problem setup in terms of performance and uses much less annotation effort. In this work we follow the setup of [2], but we go even further to only assume a pre-trained person detector instead of a detector for all objects present in a dataset (as was the case in [2]).

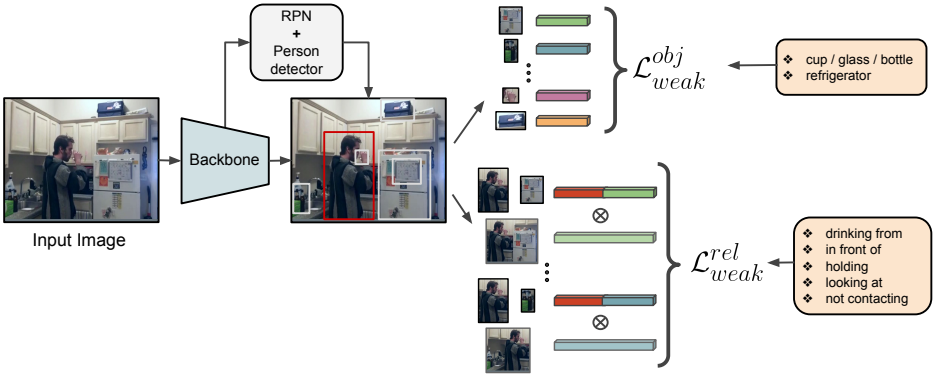

Figure 2: **Weakly-supervised obj + rel HOI**: The figure depicts our approach where the model takes in an image, uses a pretrained backbone and a Region Proposal Network (RPN) to obtain object proposals (in gray), and a person detector to detect person proposals (in red). The object features are then used by weakly-supervised object detection branch to classify object categories. For relations, the features of person and objects are concatenated and multiplied with their union features (optional) to obtain relation features, which are then used by weakly-supervised relation detection branch to classify relation categories.

3 Approach

Human Object Interaction (HOI), as the form of Visual Relation Detection (VRD) task, comprises of detecting triplets of the form `<person, predicate, object>` in an image, which requires identifying person and object bounding box along with the relation between them, *e.g.* `person is carrying a book` (shown in Figure 1). Our approach is divided into two stages. The first stage uses a pretrained object detector, based on Faster R-CNN [35], to extract person and class-agnostic object region proposals, and their RoI features from an image. The second stage forms relation candidates by pairing person with object proposal regions and applying *weak-supervision* object and relation losses to train the object and relation detector. The illustration of this process is shown in Figure 2.

Specifically, we denote an input image as $\mathbf{x}_i$ and the corresponding set of region proposals $R_i = \{\mathbf{rbox}_{i,j}\}$ and their RoI feature vectors $Z_i = \{\mathbf{z}_{i,j}\}$, where $j \in \{1,2,...,N_i\}$ and N_i denotes the total number of region proposals for image $\mathbf{x}_i$. We use the person scores obtained from pretrained object detector (or person detector) to select high scoring proposals as the person region proposals, denoted by R^P_i, and the corresponding RoI feature vectors, Z^P_i. The remaining region proposals and RoI features for class-agnostic objects are denoted as R^O_i and Z^O_i respectively; such that $R^P_i \cap R^O_i = \emptyset$ and $R^P_i \cup R^O_i = R_i$. We use a two-layered MLP to process the RoI features Z^P_i and Z^O_i to obtain person and object feature matrices respectively, *i.e.*, $\Phi^P_i \in \mathbb{R}^{N^P_i \times d}$ and $\Phi^O_i \in \mathbb{R}^{N^O_i \times d}$, where N^P_i and N^O_i is the number of person and object boxes in image i and d is the dimension of the feature vector produced by two-layered MLP. We then form the set of all possible relations as the cross product between R^P_i and R^O_i, resulting in $(N^P_i \times N^O_i)$ possible relations in total. We form the relation feature matrix by concatenating the corresponding person and object features from Φ^P_i and Φ^O_i to give $\Phi^{P\times O}_i \in \mathbb{R}^{(N^P_i \times N^O_i)\times 2d}$. Optionally, following [41], we can further enhance relation representations by extracting features from the *union* boxes obtained by taking the union of the given person and object proposal boxes; this is achieved by element-wise multiplication of *union* features with $\Phi^{P\times O}_i$.

We then use the weak object loss on the set R_i^O and weak relation loss on the set $R_i^{P\times O}$ to train the model. In particular, we build the weak object and relation classifiers, each based on WSDDN [3] framework. However, we note that our approach is flexible and can adopt any weakly-supervised loss, *e.g.*, OICR [42], PCL [43], *etc*.

3.1 Weakly-supervised object detection

Following WSDDN [3], we form two linear classifiers: f_{cls}^{obj} and f_{det}^{obj} which project a d-dimensional feature vector to $|\mathcal{C}_{obj}|+1$ output dimensions, where $\mathcal{C}_{obj}$ is the set of object classes and the additional class for background. The classifiers f_{cls}^{*} and f_{det}^{*} stand for *classification* and *detection* streams respectively, where classification stream classifies proposals into one of the target classes and detection stream focus on ensuring that not many proposals get assigned the same class [3]. Formally, we obtain:

$$S_{obj}(\mathbf{x}_i) = \text{softmax}_{cls}\left[f_{cls}^{obj}(\Phi_i^O)\right] \odot \text{softmax}_{det}\left[f_{det}^{obj}(\Phi_i^O)\right], \tag{1}$$

where softmax_{cls} is softmax over the class dimension and softmax_{det} is over the object dimension, and $S_{obj}(\mathbf{x}_i) \in \mathbb{R}^{N_i^O\times(|\mathcal{C}_{obj}|+1)}$; $\odot$ is element-wise multiplication operator.

3.2 Weakly-supervised relation detection

Similar to the above, suppose $\mathcal{C}_{rel}$ is the set of relation classes. We form two linear classifiers: f_{cls}^{rel} and f_{det}^{rel} which project $2d$-dimensional feature vectors to $|\mathcal{C}_{rel}|+1$ output dimensions.

$$S_{rel}(\mathbf{x}_i) = \text{softmax}_{cls}\left[f_{cls}^{rel}(\Phi_i^{P\times O})\right] \odot \text{softmax}_{det}\left[f_{det}^{rel}(\Phi_i^{P\times O})\right], \tag{2}$$

where softmax_{cls} is softmax over the class dimension and softmax_{det} is over the relation dimension, and $S_{rel}(\mathbf{x}_i) \in \mathbb{R}^{(N_i^P\times N_i^O)\times(|\mathcal{C}_{rel}|+1)}$.

3.3 Loss function

Suppose $\mathbf{y}_{obj}^i$ and $\mathbf{y}_{rel}^i$ are the set of image-level object and relation labels respectively for image i. We obtain predictions by aggregating the image-level classification scores for object and relation modules as $\hat{\mathbf{y}}_{obj}^i = \sum_{j=1}^{N_i^O} S_{obj}(\mathbf{x}_i)$ and $\hat{\mathbf{y}}_{rel}^i = \sum_{j=1}^{N_i^P\times N_i^O} S_{rel}(\mathbf{x}_i)$. The loss is then computed using binary cross-entropy loss function as

$$\mathcal{L}_{weak}^{obj} = -\frac{1}{|\mathcal{C}_{obj}|}\sum_{c=1}^{|\mathcal{C}_{obj}|}\left(\mathbb{1}_{[c\in\mathbf{y}_{obj}^i]}\log\hat{\mathbf{y}}_{obj}^i + \mathbb{1}_{[c\notin\mathbf{y}_{obj}^i]}\log\left(1-\hat{\mathbf{y}}_{obj}^i\right)\right) \tag{3}$$

$$\mathcal{L}_{weak}^{rel} = -\frac{1}{|\mathcal{C}_{rel}|}\sum_{c=1}^{|\mathcal{C}_{rel}|}\left(\mathbb{1}_{[c\in\mathbf{y}_{rel}^i]}\log\hat{\mathbf{y}}_{rel}^i + \mathbb{1}_{[c\notin\mathbf{y}_{rel}^i]}\log\left(1-\hat{\mathbf{y}}_{rel}^i\right)\right) \tag{4}$$

We add the above two losses to obtain the final loss, $\mathcal{L} = \mathcal{L}_{weak}^{obj} + \mathcal{L}_{weak}^{rel}$, and we jointly train both the object and relation classification modules with SGD optimizer.

4 Experiments

We present the results of our approach on two different datasets. For each dataset we provide implementation details regarding the problem setup, models, and evaluation metrics.

4.1 HICO-DET

The Humans Interacting with Common Objects (HICO-DET) dataset [5] contains 600 human-centric relation classes (*e.g.*, `wash`, `ride`, `eat`, *etc.*) and 80 object categories similar to MS-COCO [23]. The train set contains roughly $\sim 39K$ images and test set $\sim 9K$ images. We sample 15% of the train set as the validation set to be consistent with [2].

Inference Details. In inference we apply same steps as in training (Section 3) up to the loss computation. We use the relation score for scoring relation predictions. Further details are given in Section B.2 of Supplemental Material.

Evaluation metrics. We follow Baldassarre et al. [2] and use 11-point interpolated mean Average Precision (mAP) over 600 HOI classes. A predicted triplet is correct with respect to a ground-truth triplet if the following criteria are met: (1) subject, relation, and object categories match, (2) subject boxes have IoU > 0.5, (3) object boxes have IoU > 0.5, and (4) the ground-truth triplet didn't match with any previous detected triplet. In addition, we also report results using the evaluation metric from the SOTA work in HOI [40], which differs from Baldassarre et al. [2] in the following ways: (1) For an image in [2], the false positives for a predicted HOI triplet `<sub,pred,obj>` are ignored when the HOI triplet is not present in the ground truth of the image, leading to an optimistic measure, (2) and [2] uses all HOI predictions while [40] uses only top 100 predictions sorted by their scores.

Implementation Details. We use Faster R-CNN from `detectron2` [48] to remain consistent with [2]. The model is trained on MS-COCO dataset [23] and the object features are given by a feature pyramid network with ResNeXt-101 [51] backbone. We use score threshold of 0.5 in line with [2] to filter out proposals. We divide the set of obtained proposals into persons and non-person set since we assume a person detector (see Section 3), and we form relation pairs by taking cross-product between the two sets. For each proposal, we obtain a feature map of dimension $256 \times 7 \times 7$, which we pass through an `AvgPool2d` layer with kernel size 2 and padding 1, and flatten it to obtain a feature vector of 4096 dimension, which is then passed through an MLP comprising of two layers - `[Linear(4096, 2048), ReLU(), Linear(2048, 1024)]` to obtain object features. The object features are used by weakly-supervised object detection branch to predict objects. For relations the object features are concatenated together according to relation pairs to form relation features, which are then used by weakly-supervised relation detection branch to predict relations. For reproducibility, the code and with pre-trained models can be found at: https://github.com/ubc-vision/SimpleWeakHOI.

Results. Table 1 illustrates the results. Despite using weaker supervision[2], we obtain better performance as compared to [2]. Moreover, unlike [2], we do not use spatial features of objects and relation which are a function of relative object bounding box coordinates, size, and angle between the centers of bounding boxes involved in a relation. Note, we do not use frequency prior on relations here and compare with the uniform prior result of WS-VRD [2].

[2]Unlike [2], we do not assume existence of a pre-trained object detector, nor utilize its object classification scores. Instead, we only rely on person-detector and generic proposal mechanism.

Table 1: **Weakly-supervised HOI results on HICO-DET**. Performances for relevant prior works, including fully-supervised and weakly-supervised are from [2] and [40].

	Evaluation method of [2]			Evaluation method of [40]		
Method	Full (600)	Rare (138)	Non-rare (462)	Full (600)	Rare (138)	Non-rare (462)
Fully-supervised						
Chao [5]	7.81	5.37	8.54	-	-	-
InteractNet [12]	9.94	7.16	10.77	-	-	-
GPNN [33]	13.11	9.34	14.23	-	-	-
iCAN [9]	14.84	10.45	16.15	-	-	-
Analogies [30]	19.40	14.60	20.90	-	-	-
VSGNet [45]	-	-	-	19.80	16.05	20.91
FCMNet [26]	-	-	-	20.41	17.34	21.56
VCL [14]	-	-	-	23.63	17.21	25.55
ConsNet [27]	-	-	-	24.39	17.10	26.56
DRG [10]	-	-	-	24.53	19.47	26.04
UnionDet [17]	-	-	-	17.58	11.72	19.33
Wang et al. [47]	-	-	-	19.56	12.79	21.58
PPDM [22]	-	-	-	21.73	13.78	24.10
QPIC [40]	-	-	-	29.90	23.92	31.69
Ours	44.29	46.90	43.51	23.78	17.85	25.55
Baselines						
Most frequent obj and rel	0	0	0	0	0	0
Sampling from obj and rel distribution	0	0	0	0	0	0
Weakly-supervised rel						
WS-VRD [2]	24.25	20.23	25.45	-	-	-
Weakly-supervised obj+rel						
Ours	**28.77**	**24.64**	**30.00**	13.40	7.53	15.15

We include simple baselines as the lower-bound performance for our model, i.e., 1) most frequent object and relation which are `bicycle` and `hold` respectively, and (2) sampling from object and relation distribution derived from train set. We also report fully-supervised version of our model placing an upper-bound on its performance, which is able to perform comparably to the SOTA approaches despite being simplistic.

Ablation. In order to check whether the multi-task loss over objects and relations is effective, we train an additional model but with the object loss removed. Since we need object class predictions for training and evaluation for the model, we use all 80 object detectors from `detectron`'s COCO detector. We refer to this model as "rel weak loss only w/ COCO detector" (first row of Table 2). The corresponding object detection performance is 33.7 mAP @IoU 0.5. This model mimics the setup used by Baldassarre et al. [2], but we note that the performance is not at par with their reported performance since they use extra features and explanatory modules.

On the other hand, our weakly-supervised model mentioned in Table 1 uses both the weak

Table 2: **Multi-task loss ablation on HICO-DET**. We compare models trained with weak relation loss (1st row), and our proposed model trained with both weak object and relation loss (3rd row), and the model (in 2nd row) using relation detector from 1st row and object detector from 3rd row.

Method	Full (600)	Rare (138)	Non-rare (462)	Object Detection (in mAP)
rel weak loss only w/ COCO detector	22.71	16.00	24.72	33.7
rel weak loss only w/ trained detector	24.80	23.30	25.24	**38.5**
obj + rel weak loss w/ trained detector	**28.77**	**24.64**	**30.00**	**38.5**

object and relation loss, and is referred to as "obj + rel weak loss w/ trained detector" (third row of Table 2). The corresponding object detection performance is 38.5 mAP @IoU 0.5. We form another model to test the effectiveness of our multi-task loss (object and relation), where we use the relation branch from "rel weak loss only w/ COCO detector" and combine it with the object branch from "obj + rel weak loss w/ trained detector". This ensures that the difference between the second and third row is only due to the performance of the relation detector and we refer to this model as "rel weak loss only w/ trained detector" (second row of Table 2). We observe that controlling for the object detection performance, our multi-task loss outperforms the relation only loss.

4.2 Action Genome

Action Genome [15] is a scene graph data annotated over crowd-acted videos of Charades dataset [38]. Action Genome consists of 35 object categories and 25 relations and has a total of 9,848 videos, of which 7,985 belong in the train set and the remaining 1,863 to the test set, consistent with [15]. We randomly sub-sample 400 videos from the train set as the validation set. We use the annotated frames provided by [15] as not all frames in the videos are annotated. We filter out frames without person or object annotations because they would not yield a scene graph and are unusable for training or evaluation.

Inference Details. Again, in inference we apply same steps as in training (Section 3). The relation predictions are scored as a product of object scores involved and the relation score itself. Here we also use frequency prior to modulate final predictions as done in prior related works [15, 42]. Further details are given in Section B.2 of Supplemental Material.

Evaluation Metrics. We adopt standard metrics and evaluation procedures from Scene Graph Generation literature. In doing so we adopt (`SGGen`) condition where the model is neither provided with ground-truth object labels nor the bounding boxes, rather the model takes an image, predicts bounding boxes, object categories and predicate labels [15, 28]. To remain consistent with [15], we evaluate using Recall@20 and Recall@50 metrics.

Implementation Details. We first train an object detector using a Faster R-CNN architecture with feature pyramid and ResNeXt-101 [51] backbone on the Action Genome dataset. We use the RPN and RoI head of the object detector to obtain object proposals and their RoI features of 4096-dimensions. Importantly, we make no use of object classification scores (apart for a person class). During training, we remove object proposals with objectness score less than 0.5 and pick the proposal with maximum person score as the person, forming relation pairs with the person and rest of the object proposals.

Table 3: **Weakly-supervised HOI on Action Genome.** The prior work results are from [15].

Row	Method	R@20	R@50
(1)	VRD [28]	10.28	10.94
(2)	Freq Prior [56]	24.03	24.87
(3)	IMP [52]	23.88	25.52
(4)	MSDN [20]	24.00	25.64
(5)	Graph R-CNN [53]	24.12	25.77
(6)	RelDN [59]	25.00	26.21
	Baselines		
(7)	Most frequent object and relation	0.0	0.0
(8)	Sampling from obj and rel distribution	1.87	3.4
	Fully-supervised		
(9)	Ours	27.93	30.42
	Weakly-supervised obj + rel		
(10)	Ours w/o Freq prior	18.02	19.55
(11)	Ours	23.21	27.24
	Weakly-supervised obj + rel + Kinetics transfer		
(12)	Ours	24.42	28.96

We take the obtained RoI features from these boxes and pass them through a two-layered MLP (similar to the previous Section 4.1) to obtain object features. We use the features in weakly-supervised object detection branch to predict objects. For relations, we form relation features by concatenating the object features according to relation pairs, and multiplying them by the (optional) union features (depicted in Figure 2) inspired by [41]. The obtained relation features are used in weakly supervised relation detection branch to predict relations.

We form union features for each person-object pair. Suppose a candidate person-object pair has bounding box coordinates $\boldsymbol{b}_{subj}$ and $\boldsymbol{b}_{obj}$ for subject and object. We form a union bounding box and extract the RoI feature, *i.e.*, $\texttt{RoIAlign}\big(\boldsymbol{b}_{subj} \cup \boldsymbol{b}_{obj}\big)$, we then add positional encoding of union bounding to this representation. Please see [41] for further details.

Results. The results are in Table 3. Since no weakly-supervised methods exist that report performance on this dataset, we both form our own simple baselines and compare to state-of-the-art fully-supervised scene graph generation variants. Specifically, we provide results of two *simple* baselines: (7) most frequent object and relation, which are `table` and `hold` respectively, and (8) sampling from object and relation distribution, which are drawn using their occurrences in the train set. The main purpose is to observe that the task is not trivial.

Next we compare our result in both fully-supervised and weakly-supervised object and relation setup, to state-of-the-art fully-supervised baselines. Note, unless stated otherwise, all our model variants in Table 3 leverage the frequency prior introduced in [56]. Based on the results in Table 3, one can make a number of observations: (1) our simple multi-task model trained in fully-supervised manner (9) is competitive to more sophisticated models (1) - (6); (2) our weakly-supervised variant (11) sees only ~ 4.7 drop in recall compared to fully-supervised counterpart (9); (3) our weakly-supervised variant performs nearly on par with state-of-the-art scene graph generation methods in R@20 – < 1 drop in recall in most cases (1) - (5) and even outperforms *all* fully-supervised methods in R@50. Similar to [56], (4) we observe that frequency prior is important (10) vs. (11). Also Table 4 shows that multi-task loss is better than

Table 4: **Multi-task loss ablation on Action Genome dataset**. Similar to Table 2, we compare multi-task loss with relation only loss. We report performances for three categories of classes separately - `attention`, `spatial` and `contact` [15]. In addition to recall (R@20) we report mean-recall (mR@20) which is mean over individual class recalls. Both the configurations are evaluated using same pretrained detector (mAP@IoU 0.5 is 26.67)

Method	attention		spatial		contact	
	R@20	mR@20	R@20	mR@20	R@20	mR@20
rel weak loss only	32.87	32.24	40.22	36.66	**41.06**	21.21
obj + rel weak loss	**34.34**	**33.45**	**40.52**	**37.01**	40.93	**23.66**

relation only loss, drawing the same conclusion as Table 2.

Transfer learning. Besides weak-supervision, transfer learning [4, 25, 39, 49, 54] is another common strategy for efficient learning, where knowledge from the *source* domain is utilized for a *target* domain in a data-efficient manner. We conduct a small-scale experiment to see if transfer learning can further improve the performance of the above simple baseline. One unique property of Action Genome [15] dataset is that the frames are taken from Charades videos [38] which opens up the possibility for use of video-based features. We use an off-the-shelf action classification model - SlowFast 8×8 [8] with ResNet-50 as the backbone which was pretrained on Kinetics-400 dataset [16]. For every frame in Action Genome dataset, we extract a clip of 2.13 seconds around the frame in question, and compute 400-dimensional Kinetics logits. We use relation pairs in the frame, where for each relation pair, we only consider the union bounding box region in the clip for computing the logits, *i.e.*, we clip the video using the union bounding box region of the relation pair and compute the logits which are specific to the relation pair, and therefore for all relation pairs in the frame we get a 400-dimensional vector of action predictions.

We incorporate the obtained logits into the model by concatenating 400-dimensional logit to the corresponding relation features before passing it through the weakly-supervised relation detection branch. The results are shown in row (12) of Table 3 where the use of video-based features leads to improvement in performance of our weakly-supervised model. In particular, we argue that Kinetics [16] dataset is ImageNet [7] equivalent for videos and does not assume any extra information, this strategy yields models that are on-par with the prior fully-supervised works in terms of performance. Similar strategies can be adopted to HICO-DET by leveraging pre-trained still image action recognition model *e.g.*, [11, 24, 36, 60].

Qualitative results. Qualitative results for both HICO-DET and Action Genome datasets, as well as different variants of our model, can be found in the Supplemental Material.

5 Discussion and conclusions

In this paper we propose a simple, but surprisingly strong, baseline for weakly-supervised Visual Relation Detection (VRD) and scene-graph generation. Despite reliance on weaker, but in our opinion more natural, form of supervision, our approach is more accurate. It sets a new state-of-the-art in weakly-supervised VRD and performs nearly on par with fully-supervised counterparts in scene graph prediction. Following the recent trend of carefully analyzing and re-assessing performance of simple architectures on complex tasks (e.g., human pose estimation [50], audio-visual dialog [37], *etc.*), we similarly aim to provide a simple and strong building block for future research in VRD.

References

[1] Aditya Arun, CV Jawahar, and M Pawan Kumar. Dissimilarity coefficient based weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9432–9441, 2019.

[2] Federico Baldassarre, Kevin Smith, Josephine Sullivan, and Hossein Azizpour. Explanation-based weakly-supervised learning of visual relations with graph networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 612–630. Springer, 2020.

[3] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2846–2854, 2016.

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017.

[5] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 381–389. IEEE, 2018.

[6] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3076–3086, 2017.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6202–6211, 2019.

[9] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[10] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *European Conference on Computer Vision*, pages 696–712. Springer, 2020.

[11] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with r* cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1080–1088, 2015.

[12] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8359–8367, 2018.

[13] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

[14] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *European Conference on Computer Vision*, pages 584–600. Springer, 2020.

[15] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10236–10247, 2020.

[16] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[17] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *European Conference on Computer Vision*, pages 498–514. Springer, 2020.

[18] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(1):14–29, 2015.

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73, 2017.

[20] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1261–1270, 2017.

[21] Xiaodan Liang, Lisa Lee, and Eric P Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 848–857, 2017.

[22] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–490, 2020.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[24] Lu Liu, Robby T Tan, and Shaodi You. Loss guided activation for action recognition in still images. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 152–167. Springer, 2018.

[25] Lu Liu, William Hamilton, Guodong Long, Jing Jiang, and Hugo Larochelle. A universal representation transformer layer for few-shot image classification. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[26] Yang Liu, Qingchao Chen, and Andrew Zisserman. Amplifying key cues for human-object-interaction detection. In *European Conference on Computer Vision*, pages 248–265. Springer, 2020.

[27] Ye Liu, Junsong Yuan, and Chang Wen Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4235–4243, 2020.

[28] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 852–869. Springer, 2016.

[29] Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Weakly-supervised learning of visual relations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5179–5188, 2017.

[30] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting unseen visual relations using analogies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1981–1990, 2019.

[31] Bryan A Plummer, Arun Mallya, Christopher M Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. Phrase localization and visual relationship detection with comprehensive image-language cues. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1928–1937, 2017.

[32] Alessandro Prest, Cordelia Schmid, and Vittorio Ferrari. Weakly supervised learning of interactions between humans and objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(3):601–614, 2011.

[33] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–417, 2018.

[34] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 91–99, 2015.

[36] Marjaneh Safaei and Hassan Foroosh. Still image action recognition by predicting spatial-temporal pixel evolution. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 111–120. IEEE, 2019.

[37] Idan Schwartz, Alexander Schwing, and Tamir Hazan. A simple baseline for audio-visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[38] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 510–526. Springer, 2016.

[39] Krishna Kumar Singh, Santosh Divvala, Ali Farhadi, and Yong Jae Lee. Dock: Detecting objects by transferring common-sense knowledge. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 492–508, 2018.

[40] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10419, 2021.

[41] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6619–6628, 2019.

[42] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2843–2851, 2017.

[43] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(1):176–191, 2018.

[44] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 104(2):154–171, 2013.

[45] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13617–13626, 2020.

[46] Jiajie Wang, Jiangchao Yao, Ya Zhang, and Rui Zhang. Collaborative learning for weakly supervised object detection. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

[47] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4116–4125, 2020.

[48] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[49] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(9):2251–2265, 2018.

[50] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[51] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1492–1500, 2017.

[52] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5419, 2017.

[53] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 670–685, 2018.

[54] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3712–3722, 2018.

[55] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Weakly supervised visual semantic parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3736–3745, 2020.

[56] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5831–5840, 2018.

[57] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4233–4241, 2017.

[58] Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed Elgammal. Relationship proposal networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5678–5686, 2017.

[59] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11535–11543, 2019.

[60] Zhichen Zhao, Huimin Ma, and Shaodi You. Single image action recognition using semantic body part actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3391–3399, 2017.

[61] Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 843–851, 2019.

[62] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[63] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian Reid. Towards context-aware interaction recognition for visual relationship detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 589–598, 2017.