Improving Human-Object Interaction Detection via Phrase Learning and Label Composition

Zhimin Li[1*]   **Cheng Zou**[2*✉]   **Yu Zhao**[3]   **Boxun Li**[3]   **Sheng Zhong**[1✉]

[1] National Key Laboratory of Science and Technology on Multispectral Information Processing, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[2] Ant Group
[3] Megvii Technology
{lizm, zhongsheng}@hust.edu.cn   wuyou.zc@antgroup.com   {zhaoyu03, liboxun}@megvii.com

Abstract

Human-Object Interaction (HOI) detection is a fundamental task in high-level human-centric scene understanding. We propose PhraseHOI, containing a HOI branch and a novel phrase branch, to leverage language prior and improve relation expression. Specifically, the phrase branch is supervised by semantic embeddings, whose ground truths are automatically converted from the original HOI annotations without extra human efforts. Meanwhile, a novel label composition method is proposed to deal with the long-tailed problem in HOI, which composites novel phrase labels by semantic neighbors. Further, to optimize the phrase branch, a loss composed of a distilling loss and a balanced triplet loss is proposed. Extensive experiments are conducted to prove the effectiveness of the proposed PhraseHOI, which achieves significant improvement over the baseline and surpasses previous state-of-the-art methods on Full and NonRare on the challenging HICO-DET benchmark.

1   Introduction

Human is the main focus in visual world(Gupta, Hoffman, and Malik 2016). Figuring out the relationships between human and its context is of great importance. Human-Object Interaction (HOI) is to detect interactive relation between human and the surroundings. HOI detection plays an important role in high-level human-centric scene understanding, and has attracted considerable research interest recently (Ulutan and Iftekhar 2020; Kim et al. 2020; Liao et al. 2020; Wang et al. 2020; Kim et al. 2020; Liu, Chen, and Zisserman 2020; Bansal et al. 2020; Gao, Zou, and Huang 2018), resulting in significant improvement (Gao et al. 2020; Hou et al. 2020; Li et al. 2020; Zou et al. 2021).

The task of HOI detection aims at localizing and recognizing relationships between humans and objects. Fig. 1(a) illustrates an example of HOI, which inputs an image and outputs human bounding boxes, object bounding boxes and the interaction categories between them.
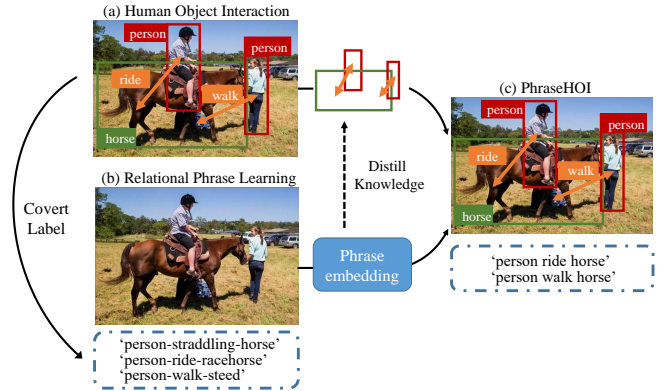

Figure 1: (a) HOI task, which inputs an image and outputs human bounding boxes, object bounding boxes and the interactions between them. (b) Relational phrase learning task, which inputs an image and outputs several natural language phrases to describe the interactions between humans and objects. (c) Proposed PhraseHOI, which inputs an image and outputs all of the above.

It's hard to obtain labeled data for HOI due to the various interactions between humans and objects in reality. So it's natural to think about using data from other sources as an alternative way to improve the performance of HOI detection. In this paper, we propose a novel relational phrase learning task derived from HOI to utilize language knowledge. Fig. 1(b) illustrates an example of relational phrase learning. It inputs an image and outputs several natural language phrases to describe the interactions between humans and objects. Here we call these phrases 'relational phrases', and the ground truths for them are automatically converted from the original HOI annotations without extra human efforts.

Once staring at Fig. 1(a) and Fig. 1(b) together, we find that these two tasks both model the interactions between humans and objects, although one is from visual modality and the other is from language modality. We hypothesize that these two tasks share a common part in there knowledge space, and joint optimization may help to distill knowledge from the relational phrase learning task to HOI task, which

will improve the performance of HOI detection. Specifically, influenced by the task of cross-modal knowledge distillation (Gupta, Hoffman, and Malik 2016), our knowledge comes from the teacher, a pretraind word2vec model, then to teach the HOI model.

Also, HOI detection is confronted with a long-tailed problem due to the large enumeration space of verb-object pairs. This is really a challenge because there exists very few samples for some rare HOI categories. To recognize something unknown, we observe that human beings tend to recall information from those related or similar things they have already known. For instance, the scene of 'person ride giraffe' is rare, however human have learned the prior knowledge of 'person ride horse' and 'person watch giraffe', then we could quickly compose 'person ride giraffe' with the prior knowledge and remember it. Inspired by this, we construct semantic neighborhood set for each word and then propose a novel label composition method to extend label space. Specifically, we composite new labels for a given relational phrase by replacing a verb word or an object word according to their semantic neighbors. This composition method is quite different from previous work (Bansal et al. 2020; Hou et al. 2020), all of which do augmentation in input feature space and need to break the pipeline, while ours is in the label space.

In this paper, we propose a multi-task architecture named PhraseHOI to jointly optimize the related tasks, i.e. HOI and relational phrase learning, resulting in significant performance improvement on HOI. It consists of a transformer backbone and two parallel branches. One is HOI branch, and its design follows (Zou et al. 2021). The other is phrase branch, it outputs an embedding which can be translated into a relational phrase through Look-Up-Table (LUT). Fig. 1(c) illustrates the concept of the proposed method, which inputs an image and outputs human bounding boxes, object bounding boxes, interaction categories, and relational phrases. To optimize the phrase branch, a distilling loss along with triplet loss is proposed, to globally keep the feature distribution of the original language space, and to locally make semantically similar embeddings discriminative to each other.

The contributions are summarized as follows,

- As far as we know, we are the first to creatively use HOI annotations as phrases. And based on this, we propose a novel relational phrase learning task derived from HOI to describe interactions between humans and objects with natural language phrases.

- We propose PhraseHOI to jointly optimize HOI detection and relational phrase learning to improve the performance of HOI.

- We propose a novel label composition method to address the long-tailed problem in HOI, which composites new samples by semantic neighbors in label space.

- Our method gets significant improvement compared to our baseline, set state-of-the-art on Full and NonRare on the challenging HICO-DET benchmark.

2 Related Work

External Knowledge. Using knowledge of different types to improve HOI detection is becoming popular, because these extra knowledge brings in richer information. In (Li et al. 2020, 2019; Shen et al. 2018; Wan et al. 2019; Zhou and Chi 2019; Zhong et al. 2020; Fang et al. 2018), human pose or body part is introduced to pay more attention to local area rather than treating the whole human area uniformly. In (Xu et al. 2019; Gao et al. 2020; Liu, Chen, and Zisserman 2020; Kim et al. 2020; Bansal et al. 2020), word embedding or language prior is introduced, e.g. Xu (Xu et al. 2019) used GloVe (Pennington, Socher, and Manning 2014) to generate word embedding as part of graph node representation. In this work, we use language prior to enhance HOI detection, but unlike other methods, the language embedding in our work is the output and it is the target for the model to learn, rather than the input feature to the model. This makes the model flexible and can be regarded as a meta-architecture for other HOI detectors.

Long-tailed Problem. HOI detection is facing a long-tailed distribution problem because it is difficult to collect labeled training data for all HOI categories, due to the large enumeration space of verb-object pairs. Shen (Shen et al. 2018) may be the first one to introduce zero-shot learning approach to tackle the long-tailed problem in HOI by disentangling reasoning on verbs and objects. Bansal (Bansal et al. 2020) incorporates the idea that human interacting with functionally similar objects look similar via a functional generalization module, resulting in feature augmentation during training according to the semantic representation similarities of objects. VCL (Hou et al. 2020) addresses the long-tailed issues via visual composition learning, which composes new interaction samples in the visual feature space by decomposing an HOI representation into object and verb specific features. These two methods are doing augmentation in input feature space, while in this work, we propose a novel label composition method to augment in the label space, which is quite different from (Bansal et al. 2020; Hou et al. 2020).

Context Modeling. It's vitally important to model the context information properly for HOI detection. In (Gao, Zou, and Huang 2018; Li et al. 2019; Wang et al. 2019; Ulutan and Iftekhar 2020), a two-channel binary image representation is used to encode the spatial context. In (Liu, Chen, and Zisserman 2020), image scene class information is introduced to encode global visual context. DRG (Gao et al. 2020) realized the problem that most methods ignore the contextual information from other interactions in the scene and predict the current interaction between each human-object pair in isolation, so a dual graph is proposed to enable knowledge transfer among objects. CHG (Wang, Zheng, and Yingbiao 2020) found that relations between homogeneous entities and heterogeneous entities should not be equally the same, so a contextual heterogeneous graph is built to model the intra-class context and inter-class context, which is more elaborate than (Zhou and Chi 2019). Most recently, (Zou et al. 2021; Tamura, Ohashi, and Yoshinaga 2021; Chen et al. 2021; Zhang et al. 2021) reasons about the interactions between humans and objects from global image context with
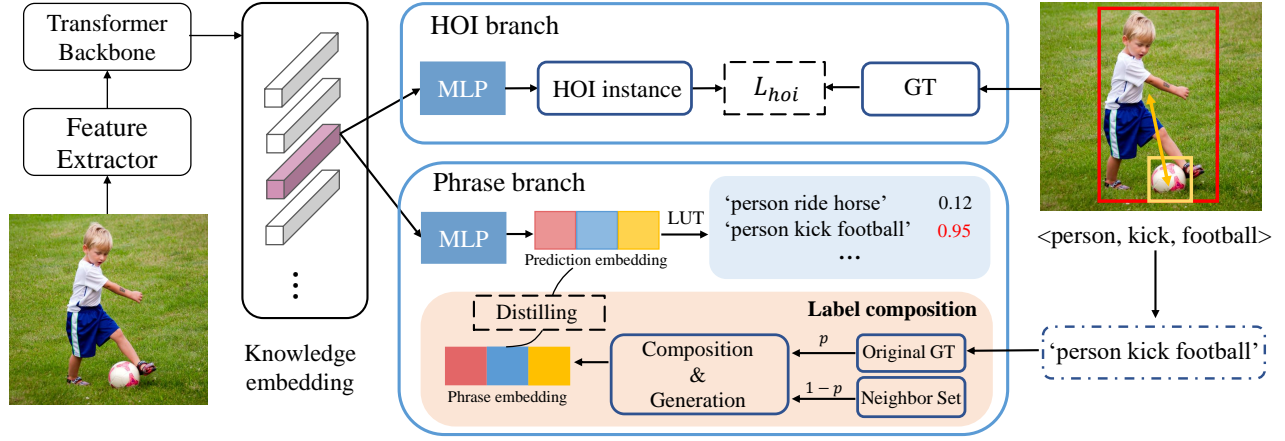
Figure 2: Overall architecture. The proposed PhraseHOI has a transformer backbone and two parallel branches, one is HOI branch and the other is phrase branch. The training and inference of HOI branch follows (Zou et al. 2021). The phrase branch first predicts phrase embedding and then translates it into a relational phrase through Look-Up-Tab (LUT). In training, the original HOI triplet is first converted to a relational phrase, and then sent to a label composition module to generate new phrases, after that, the phrase is encoded to a phrase embedding as GT to supervise the prediction embedding.

transformer. In this work, we introduce transformer to model the context information following (Zou et al. 2021) due to its large receptive field and simple architecture.

3 Methodology

3.1 Overview

In this section, we first propose a multi-task architecture named PhraseHOI to jointly optimize the related tasks, i.e. HOI and relational phrase learning, in a unified network. Second, to learn the relational phrase task, we creatively generate ground truths from original HOI annotations. Then, a novel label composition method is proposed to handle the long-tailed problem in HOI. After that, it's the phrase embedding learning loss. Finally, we introduce two available post processing methods.

3.2 Network Architecture

The proposed PhraseHOI has a transformer backbone and two parallel branches, one is HOI branch and the other is phrase branch. The HOI branch outputs HOI instance, i.e. a quintuple of (human class, interaction class, object class, human box, object box) and the phrase branch outputs a phrase embedding which can be translated to a relational phrase through Look-Up-Table (LUT).

Transformer Backbone. The network design follows (Zou et al. 2021) because transformer provides large receptive field and rich context information, both locally and globally, which is important for HOI detection. Given an input image, the transformer backbone outputs N feature embeddings (denoted as knowledge embedding in Fig. 2), which have encoded the relationship knowledge between humans and objects implicitly.

HOI Branch. This branch is attached to the knowledge embedding, and consists of five MLP heads which directly predict the HOI instance, i.e. a quintuple of (human class, interaction class, object class, human box, object box). The training loss and training target for this branch follows (Zou et al. 2021).

Phrase Branch. The phrase branch lies in parallel to the HOI branch, and both of them are attached to the same knowledge embedding, implying that these two branches share the same knowledge space. As a result, there is always a phrase embedding counterpart associated with an HOI instance in the output. After a three-layer MLP, the knowledge embedding is converted to a phrase embedding, which is a 900-d vector, and can be translated to a relational phrase through LUT. In training, the original relational phrase is first sent to a label composition module to composite a new phrase. Then the composed phrase is sent to an embedding generation module to generate a 900-d phrase embedding, which is the training target for the phrase branch. The model is jointly optimized and the ground truth for training the phrase branch is detailed in the following section. Fig. 2 illustrates the overall architecture.

3.3 Relational Phrase Generation

To train the novel phrase branch, we creatively construct a relational phrase target from a given original HOI annotation. Specifically, we directly use the 'word' of human, verb and object to construct a three-word phrase, e.g. the phrase 'human ride horse' can be derived from the original HOI triplet (human, ride, horse). After that, each of the words is sent to a pretrained word embedding model, e.g. word2vec (Mikolov et al. 2013), to get an embedding. Finally, the three embeddings are concatenated in 'human-verb-object' order to form a phrase embedding, and this is the ground truth for the phrase branch to learn.

It's noted that some compound words can not be found in word embedding model, e.g. 'dining table'. In this case we split it at first and then get the mean embedding of them.
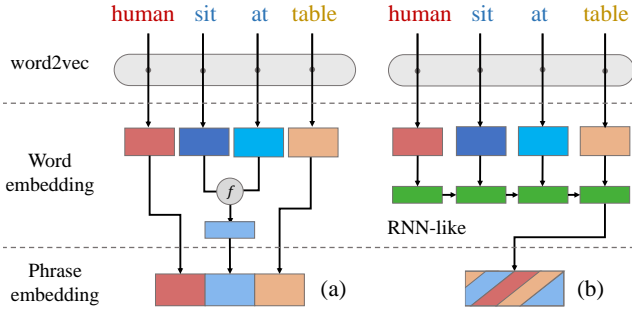
Figure 3: Illustration of two methods for phrase embedding generation. (a) the concatenated method. (b) the language model encoded method.

Fig. 3(a) illustrates this procedure. And there is another reasonable embedding generation method shown in Fig. 3(b), feeding all the words to a time serial model to get the output hidden embedding, which will be discussed in experiments.

3.4 Label Composition

HOI detection is facing a long-tailed problem due to the lack of training data on rare HOI categories. Recent studies (Bansal et al. 2020; Hou et al. 2020) indicates that augmentation strategies do help. According to (Lu et al. 2016), there is a good property in language feature space that semantically similar objects are close to each other. Based on this, we propose a novel label composition method to tackle the long-tailed problem in HOI, as shown in Fig. 4.

It is noteworthy that the proposed method is quite different from (Hou et al. 2020) and (Bansal et al. 2020). Both of them do augmentation in input feature space, which usually needs a multi-stage pipeline, while ours is in label space, nearly all kinds of pipelines fit, two-stage, one-stage, and end-to-end.

To make it simple, let's take HICO-DET dataset which contains 117 different verbs and 80 different objects as an example. For each verb word, we find its top K nearest neighbors in the whole space, and only keep those words whose similarities are larger than a given similarity threshold t_{sim}. Thus, each verb word will have its own neighborhood set. For example, given query word 'kiss' and similarity threshold $t_{sim} = 0.7$, we get its neighborhood set {('kisses', 0.81), ('smooch', 0.76), ('kissing', 0.71)}, where the float number in the bracket indicates the similarity with query word 'kiss'. It's the same with each object word.

In training, given a ground truth, e.g. 'human kiss horse', then for verb 'kiss', we sample from its neighborhood set {'kisses', 'smooch', 'kissing'} at a probability of p_v and $1 - p_v$ to keep itself, for object 'horse', we sample from its neighborhood set {'horses', 'racehorse', 'stallion', 'steed'} at a probability of p_o and $1 - p_o$ to keep unchanged. In this way, the original one ground truth 'human kiss horse' is expanded to 20 different samples, which dramatically increase the label diversity and is proved to be effective.

This augmentation makes sense because the neighborhood set for each query word is carefully chosen by their
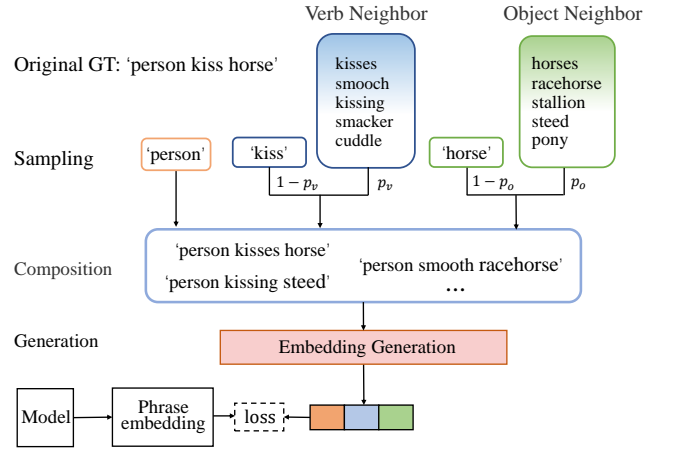

Figure 4: Illustration of the label composition method. Given a GT phrase in the format of 'human verb object', we sample from the verb word's neighborhood set and the object word's neighborhood set to composite new phrases.

similarities in the whole space, i.e. only top closest neighbors are taken into consideration. In other words, a word is replaced by its semantically similar neighbors during new phrase composition.

3.5 Distill with Triplet Loss

We design the following loss to make sure that the multi-branch can be jointly trained:

$$\mathcal{L}_{total} = \mathcal{L}_{hoi} + \alpha * \mathcal{L}_{phrase} \quad (1)$$

where $\mathcal{L}_{hoi}$ is the loss for HOI branch, which consists of the bounding box regression loss for human and object boxes, the classification loss for human, object and interaction categories. The implementation details of $\mathcal{L}_{hoi}$ follows (Zou et al. 2021). $\mathcal{L}_{phrase}$ implies the phrase embedding learning loss for phrase branch, and α is used to balance these two tasks in training.

Intuitively, $\mathcal{L}_{phrase}$ should be a distilling loss, since the relational phrase learning can be interpreted as a knowledge distilling procedure, and it aims to predict a phrase embedding which is close to the given ground truth. However, this is not enough. Distilling only constrains the relationship between predictions and ground truth, it is trying to pull the faraway predictions close to the ground truth, but it does not consider the relationship between predictions. Specifically, the predictions belonging to different class clusters may be close to each other. To solve this, a loss to constrain the relationship between predictions is required. Here we choose the classical triplet loss, which may not be the best, but it works well. Thus $\mathcal{L}_{phrase}$ is defined as:

$$\mathcal{L}_{phrase} = \mathcal{L}_{distilling} + \beta * \mathcal{L}_{triplet} \quad (2)$$

where $\mathcal{L}_{distilling}$ denotes the knowledge distilling function, and we set it to L1 loss, details are shown in ablation study. β is a hyper-parameter to balance the weight between distilling loss and triplet loss. The triplet loss $\mathcal{L}_{triplet}$ is defined as:

$$\mathcal{L}_{triplet} = max((d(A, P) - d(A, N) + m), 0) \quad (3)$$

Methods	Backbone	Knowledge	Default Full↑	Rare↑	NRare↑	Known Object Full↑	Rare↑	NRare↑
Two-stage methods								
CHG (Wang, Zheng, and Yingbiao 2020)	ResNet-50		17.57	16.85	17.78	21.00	20.74	21.08
VSGNet (Ulutan and Iftekhar 2020)	ResNet152		19.80	16.05	20.91	-	-	-
FCMNet (Liu, Chen, and Zisserman 2020)	ResNet-50	✓	20.41	17.34	21.56	22.04	18.97	23.12
ACP (Kim et al. 2020)	ResNet-152	✓	20.59	15.92	21.98	-	-	-
Bansal et al. (Bansal et al. 2020)	ResNet-50-FPN		21.96	16.43	23.62	-	-	-
PD-Net (Zhong et al. 2020)	ResNet-152		20.81	15.90	22.28	24.78	18.88	26.54
PastaNet (Li et al. 2020)	ResNet-50	✓	22.65	21.17	23.09	24.53	23.00	24.99
VCL (Hou et al. 2020)	ResNet101		23.63	17.21	25.55	25.98	19.12	28.03
DRG (Gao et al. 2020)	ResNet-50-FPN		24.53	19.47	26.04	27.98	23.11	29.43
One-stage methods								
UnionDet (Bumsoo et al. 2020)	ResNet-50-FPN		17.58	11.52	19.33	19.76	14.68	21.27
IPNet (Wang et al. 2020)	Hourglass		19.56	12.79	21.58	22.05	15.77	23.92
PPDM (Liao et al. 2020)	Hourglass		21.73	13.78	24.10	24.58	16.65	26.84
AS-Net (Chen et al. 2021)	ResNet-50		28.87	**24.25**	30.25	31.74	27.07	33.14
End-to-End methods								
HOTR (Kim et al. 2021)	ResNet-50		25.10	17.34	27.42	-	-	-
QPIC (Tamura, Ohashi, and Yoshinaga 2021)	ResNet-50		29.07	21.85	31.23	31.68	24.14	33.93
HOI Transformer (Zou et al. 2021)	ResNet-50		23.46	16.91	25.41	26.15	19.24	28.22
HOI Transformer (Zou et al. 2021)	ResNet-101		26.61	19.15	28.84	29.13	20.98	31.57
Ours	ResNet-50	✓	29.29	22.03	31.46	31.97	23.99	34.36
Ours	ResNet-101	✓	**30.03**	23.48	**31.99**	**33.74**	**27.35**	**35.64**

Table 1: Comparison to state-of-the-art methods on HICO-DET test set. Knowledge means the methods use pose or language knowledge. Our method achieves the best performance on Full and NonRare categories. NRare is short for NonRare.

where A denotes the anchor point which is each output embedding itself, P denotes positive point if it belongs to the same class cluster as A, otherwise the point is regarded as negative and denoted as N. The notation d is a distance metric, and hyper-parameter m stands for the distance margin.

3.6 Inference

The proposed network can parallelly output a group of N HOI instances. We provide two different ways: 1) predict HOI instance from HOI branch, 2) predict HOI instance from phrase branch. This can be formalized as:

$$\begin{cases} C_{hoi} = softmax(O_{hoi}) \\ C_{phrase} = LUT(O_{phrase}) \end{cases} \qquad (4)$$

For HOI branch, C_{hoi} denotes the human, object and interaction confidences, which are obtained by applying $softmax$ on the corresponding output O_{hoi} of HOI branch.

For phrase branch, the output embedding O_{phrase} is sent to query its top similar embedding in a given Look-Up-Table, and C_{phrase} is their similarity. The LUT here is often a close-set dictionary, which only contains the embeddings occurred in training set. In this way, the phrase embedding can be translated to an HOI instance. For example, if the top similar embedding in LUT stands for 'person sit at dining table', then it can be translated to the HOI instance (human, sit_at, dining_table). For both branches, we use shared human and object boxes.

It's noteworthy that there is no need to inference on both branches. In the real-world application, one can remove any

one of the branches for memory or computation efficiency. We only report the result of HOI branch in the following experiments if no extra descriptions.

4 Experiments

4.1 Experimental Settings

Datasets. Experiments are conducted on V-COCO (Gupta and Malik 2015) and HICO-DET (Chao et al. 2018) benchmark. V-COCO is a subset of MS-COCO, which consists of 5,400 images in the trainval dataset and 4946 images in test set. Each human is annotated with binary labels for 29 different action categories. HICO-DET consists of 47,776 images, 38,118 images in training set and 9,658 in test set. It has 600 HOI categories over 117 interactions and 80 objects.

Evaluation Metric. Following the standard evaluation scheme, we use mean average precision (mAP) to examine the model for both datasets. An HOI detection is considered as true positive if and only if it localizes the human and object accurately, i.e. the Interaction-over-Union (IOU) between the predicted box and ground-truth is greater than 0.5, and predicts the interaction correctly.

Implementation Details. The data augmentation follows (Zou et al. 2021), but the main difference is that, when using scale augmentation, we scale the input image such that the shortest side is at least 448 and at most 640 while the longest at most 900, which is much smaller, resulting in a larger training batch size and a shorter iteration period. This difference also contributes to a little performance gain compared to original HOI Transformer.

Methods	HOI branch			Phrase branch		
	Full	Rare	NRare	Full	Rare	NRare
HOI model	23.46	16.91	25.41	-	-	-
HOI model*	24.07	17.01	26.18	-	-	-
Phr. model	-	-	-	20.09	9.96	23.12
PhraseHOI	**29.29**	**22.03**	**31.46**	24.40	14.04	27.50

(a) Effectiveness of multi-task.

t_{sim}	Full	Rare	NRare
1.0	27.36	19.78	29.63
0.8	28.47	21.78	30.47
0.7	**29.29**	**22.03**	**31.46**
0.6	27.80	20.52	29.97

(b) t_{sim} for label composition.

$loss_{phrase}$	Full	Rare	NRare
w/o	24.07	17.01	26.18
MSE	27.06	20.18	29.11
KL div.	26.52	18.04	29.05
Cosine	26.32	17.29	29.02
$\mathcal{L}_{distilling}$	28.10	18.64	30.92
$\mathcal{L}_{phrase}$	**29.29**	**22.03**	**31.46**

(c) loss for phrase learning.

Table 2: Ablations on HICO-DET. (a-c) prove the effectiveness of joint optimization, label composition and phrase learning loss respectively. * denotes the reproduction of (Zou et al. 2021), which is our baseline. NRare is short for NonRare.

Methods	Backbone	AP_{role}
Two-stage methods		
VCL (Hou et al. 2020)	R101	48.3
Zhou et al. (Zhou et al. 2020)	R50	48.9
PastaNet (Li et al. 2020)	R50	51.0
DRG (Gao et al. 2020)	R50-FPN	51.0
VSGNet (Ulutan and Iftekhar 2020)	R152	51.8
PD-Net (Zhong et al. 2020)	R152	52.6
ACP (Kim et al. 2020)	R152	53.2
One-stage methods		
UnionDet (Bumsoo et al. 2020)	R50-FPN	47.5
IPNet (Wang et al. 2020)	HG-104	51.0
AS-NET (Chen et al. 2021)	R50	53.9
End-to-End methods		
HOTR (Kim et al. 2021)	R50	55.2
QPIC (Tamura, Ohashi, and Yoshinaga 2021)	R50	58.8
HOI transformer (Zou et al. 2021)	R50	51.2
Ours	R50	57.4

Table 3: Comparison to state-of-the-art on V-COCO test set. Our method achieves the best on AP_{role}.

The experiments are conducted on two backbones, ResNet-50 and ResNet-101. We use a COCO pre-trained DETR (Carion et al. 2020) to initialize the weights. The model is trained with AdamW, and the learning rate is set to 1e-4 except that the learning rate for backbone is set to 1e-5. The batch size for ResNet-50 and ResNet-101 are set to 64 and 32 respectively, which is 4x larger than that in HOI Transformer. All the models are trained for 200 epochs with once learning rate decay at epoch 150.

The pre-trained word embedding model and language model we use are word2vec (Mikolov et al. 2013) and GPT-1 (Radford et al. 2018), respectively. The hyper-parameter α in Eq. 1 implies the loss weight of phrase and is set to 0.1, the hyper-parameter β in Eq. 2 implies the loss weight of triplet loss and is set to 10, the hyper-parameter m in Eq. 3 is set to 0.5 in experiments.

4.2 Comparison to State-of-the-Art

We report quantitative results in terms of AP on HICO-DET in Tab. 1 and AP_{role} on V-COCO in Tab. 3. For HICO-DET dataset, our method outperforms our baseline (Zou et al. 2021) in a large margin, and set state-of-the-art on Full and NonRare categories. Compared to our baseline, we achieve 5.83% gain on Full, especially 5.12% gain on Rare cate-

gories with ResNet-50. For V-COCO dataset, amazing 6.2% gain is also achieved compared to our baseline.

4.3 Ablation Study

The ablation studies are conducted with ResNet-50 backbone models, and all the models are trained for 200 epochs with once learning rate decay at epoch 150.

Phrase Learning and Multi-Task. The novel relational phrase learning task is derived from HOI, whose labels are converted from HOI annotations. In order to evaluate it, we use its output embeddings to query from LUT to get a result of HOI, which can indicate the quality of the learned embeddings, indirectly reflecting the performance.

We test on three models to prove the effectiveness, and the results are reported in Tab. 2a. The first model is our baseline, which has only HOI branch, denoted as HOI model*, the second model is relational phrase learning model, which has only phrase branch with paired (human, object) bounding boxes, denoted as Phr. model, the third model has both branches, denoted as PhraseHOI.

As shown in Tab. 2a, the Phr. model achieves 20.09% on Full and surpasses many recent works in Tab. 1, which proves that the relational phrase learning task (with only one phrase branch) has the ability to model the interaction knowledge. Our baseline is HOI model*, a reproduction of HOI Transformer, which is slightly better than the results reported in (Zou et al. 2021), probably due to the larger training batch size. Compared to the baseline, PhraseHOI achieves 5.22% gain on Full on HOI branch, which proves the hypothesis that joint optimization may help to distill knowledge from the relational phrase learning task to HOI task. Moreover, compared to Phr. model, PhraseHOI also achieves 4.31% gain on Full on phrase branch, implying that joint training of the multi-task is beneficial to both of the related tasks.

Label Composition. According to previous section, the proposed novel label composition strategy has two degrees of freedom, one is verb, and the other is object. Best result is obtained when p_v is set to 0.8 and p_o is set to 0.2 in experiments. Here we study different values for the similarity threshold t_{sim} for object and verb neighborhood set. As shown in Tab. 2(b), $t_{sim} = 1.0$ means that the neighborhood set has only itself. It can be seen that when setting to 0.8 or 0.7, dramatic improvement are acquired, which indicates that the proposed composition method, based on semantic

m	Full	Rare	NonRare
w/o	28.10	18.64	30.92
0.01	28.76	20.09	31.35
0.10	29.10	21.94	31.24
0.50	**29.29**	**22.03**	**31.46**
1.00	28.59	21.58	30.68

Table 4: Varying m for triplet loss.

Methods	Full	Rare	NonRare
Concatenation	29.29	22.03	31.46
LanguageModel	22.64	14.12	25.19

Table 5: Different embedding generation methods.

Methods	Full	Rare	NonRare
word2vec	29.29	22.03	31.46
random	21.90	13.87	24.30

Table 6: Random word embedding experiment.

similarity and manipulated in label space, is effective. However, when setting to 0.6, the performance damages slightly, which is probably caused by the low quality of the neighborhood set when the similarity threshold is getting lower.

Phrase Learning Loss. Five different losses for phrase embedding learning are studied: MSE loss, KL divergence, cosine loss, L1($\mathcal{L}_{distilling}$) and $\mathcal{L}_{phrase}$. The distilling loss used here is to constrain the distribution of predictions equal or similar to the distribution of ground truth. As can be seen in Tab. 2(c), compared with baseline, the distilling loss is effective and gets significantly better result. However, further studies showed that using distilling losses like L1 alone is not enough, because these losses only consider the relationship between predictions and ground truth. On the contrary, relationship between predictions should also be taken into consideration, especially when the ground truths are close to each other. As shown in Tab. 2(c), using $\mathcal{L}_{phrase}$ can still help improving the performance further.

At this point, it is clear that in phrase embedding learning, L1 loss plays a more global role to gather all the predictions together around their corresponding ground truth. Meanwhile, the triplet loss does some subtle tuning in local area to push the similar but category-different points away from each other. It can also be interpreted this way: the L1 loss tries to make the intra-class variance smaller, while the triplet loss helps to make the inter-class variance larger. Results for various m are shown in Tab. 4. Setting $m = 0.50$ gives gains of 1.19% AP on Full and especial 3.39% AP on Rare.

Embedding Generation. We conduct experiments to test two different embedding generation methods, one is concatenated embedding as shown in Fig 3(a), the other is language model encoded hidden embedding as shown in Fig 3(b). The concatenated embedding is the concatenation of three embeddings in 'human-verb-object' order, each embedding is queried from word2vec (Le and Mikolov 2014) by the query word. If it is a compound word, e.g. 'sit at', which can not be found in word2vec, we calculate the mean of each of its separate part.

As shown in Tab. 5, the concatenation method performs much better in our setting. Further investigations find that the similarities between language model generated embeddings are much higher than that of the concatenated embeddings, e.g. the similarity between 'human sip wine glass'

and 'human toast wine glass' is 0.96 for language model generated embeddings, while only 0.71 for the concatenated embeddings. This may be the reason why the performance of language model generated embedding is poor, because it damages the discriminative ability of knowledge space during training.

Analysis. To further explore the reason why relational phrase helps, we conduct an extra experiment by using a set of random embeddings to replace the word embeddings from word2vec. The result in Tab. 6 shows that using random word embeddings is much poorer than that of using word2vec, which implies that it is more the language prior knowledge than the continuous label encoding (compared to discrete one-hot label encoding) structure helps in the multi-task learning.

From an HOI point of view, the proposed PhrasenHOI can be interpreted as knowledge distilling or knowledge transfer from language prior, because inference can be done without the phrase branch, but the performance of HOI branch has been improved.

From the zero-shot view, the proposed PhraseHOI has the potential to do zero-shot prediction, e.g. given an output phrase embedding from phrase branch, if we replace the LUT from close-set to open-set, the output may be semantically reasonable, but this is beyond the scope of this paper.

5 Conclusion

In this paper, we propose a derived task from HOI detection, namely relational phrase learning, which encodes semantic knowledge and generates phrases to describe the interactions between humans and objects in image. Then based on the hypothesis that HOI and relational phrase learning may share a common part in their knowledge space, we propose a multi-task architecture named PhraseHOI to jointly optimize them, aiming to improve the discriminative ability and capacity of the shared knowledge space. More efforts are made on the phrase branch, e.g. embedding distilling with triplet loss, label composition to deal with long-tailed problem, which in turn boost the performance of HOI branch. As a result, PhraseHOI gets significant improvement over baseline, setting state-of-the-art on Full and NonRare on the challenging HICO-DET benchmark, which proves the effectiveness of the proposed method. We hope our method can bring new insight to the community.

References

Bansal, A.; Rambhatla, S. S.; Shrivastava, A.; and Chellappa, R. 2020. Detecting Human-Object Interactions via Functional Generalization. In *AAAI*, 10460–10469.

Bumsoo, K.; Taeho, C.; Jaewoo, K.; and Hyunwoo, J. K. 2020. UnionDet: Union-Level Detector Towards Real-Time Human-Object Interaction Detection. In *European Conference on Computer Vision*. Springer.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. *arXiv preprint arXiv:2005.12872*.

Chao, Y.-W.; Liu, Y.; Liu, X.; Zeng, H.; and Deng, J. 2018. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, 381–389. IEEE.

Chen, M.; Liao, Y.; Liu, S.; Chen, Z.; Wang, F.; and Qian, C. 2021. Reformulating hoi detection as adaptive set prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9004–9013.

Fang, H.-S.; Cao, J.; Tai, Y.-W.; and Lu, C. 2018. Pairwise body-part attention for recognizing human-object interactions. In *Proceedings of the European conference on computer vision (ECCV)*, 51–67.

Gao, C.; Xu, J.; Zou, Y.; and Huang, J.-B. 2020. Drg: Dual relation graph for human-object interaction detection. In *European Conference on Computer Vision*, 696–712. Springer.

Gao, C.; Zou, Y.; and Huang, J.-B. 2018. iCAN: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*.

Gupta, S.; Hoffman, J.; and Malik, J. 2016. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2827–2836.

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

Hou, Z.; Peng, X.; Qiao, Y.; and Tao, D. 2020. Visual compositional learning for human-object interaction detection. *arXiv preprint arXiv:2007.12407*.

Kim, B.; Lee, J.; Kang, J.; Kim, E.-S.; and Kim, H. J. 2021. HOTR: End-to-End Human-Object Interaction Detection with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 74–83.

Kim, D.-J.; Sun, X.; Choi, J.; Lin, S.; and Kweon, I. S. 2020. Detecting human-object interactions with action co-occurrence priors. *arXiv preprint arXiv:2007.08728*.

Le, Q.; and Mikolov, T. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, 1188–1196. PMLR.

Li, Y.-L.; Xu, L.; Liu, X.; Huang, X.; Xu, Y.; Wang, S.; Fang, H.-S.; Ma, Z.; Chen, M.; and Lu, C. 2020. PaStaNet: Toward Human Activity Knowledge Engine. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 382–391.

Li, Y.-L.; Zhou, S.; Huang, X.; Xu, L.; Ma, Z.; Fang, H.-S.; Wang, Y.; and Lu, C. 2019. Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3585–3594.

Liao, Y.; Liu, S.; Wang, F.; Chen, Y.; Qian, C.; and Feng, J. 2020. 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*, 482–490.

Liu, Y.; Chen, Q.; and Zisserman, A. 2020. Amplifying key cues for human-object-interaction detection. In *European Conference on Computer Vision*, 248–265. Springer.

Lu, C.; Krishna, R.; Bernstein, M.; and Fei-Fei, L. 2016. Visual relationship detection with language priors. In *European conference on computer vision*, 852–869. Springer.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training.

Shen, L.; Yeung, S.; Hoffman, J.; Mori, G.; and Fei-Fei, L. 2018. Scaling human-object interaction recognition through zero-shot learning. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1568–1576. IEEE.

Tamura, M.; Ohashi, H.; and Yoshinaga, T. 2021. QPIC: Query-Based Pairwise Human-Object Interaction Detection with Image-Wide Contextual Information. In *CVPR*.

Ulutan; and Iftekhar. 2020. VSGNet: Spatial Attention Network for Detecting Human Object Interactions Using Graph Convolutions. *arXiv: Computer Vision and Pattern Recognition*.

Wan, B.; Zhou, D.; Liu, Y.; Li, R.; and He, X. 2019. Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 9469–9478.

Wang, H.; Zheng, W.-s.; and Yingbiao, L. 2020. Contextual Heterogeneous Graph Network for Human-Object Interaction Detection. *arXiv preprint arXiv:2010.10001*.

Wang, T.; Anwer, R. M.; Khan, M. H.; Khan, F. S.; Pang, Y.; Shao, L.; and Laaksonen, J. 2019. Deep contextual attention for human-object interaction detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 5694–5702.

Wang, T.; Yang, T.; Danelljan, M.; Khan, F. S.; Zhang, X.; and Sun, J. 2020. Learning Human-Object Interaction Detection using Interaction Points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4116–4125.

Xu, B.; Wong, Y.; Li, J.; Zhao, Q.; and Kankanhalli, M. S. 2019. Learning to detect human-object interactions with

knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhang, A.; Liao, Y.; Liu, S.; Lu, M.; Wang, Y.; Gao, C.; and Li, X. 2021. Mining the Benefits of Two-stage and One-stage HOI Detection. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Zhong, X.; Ding, C.; Qu, X.; and Tao, D. 2020. Polysemy deciphering network for human-object interaction detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, 69–85. Springer.

Zhou, P.; and Chi, M. 2019. Relation Parsing Neural Network for Human-Object Interaction Detection. 843–851.

Zhou, T.; Wang, W.; Qi, S.; Ling, H.; and Shen, J. 2020. Cascaded human-object interaction recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4263–4272.

Zou, C.; Wang, B.; Hu, Y.; Liu, J.; Wu, Q.; Zhao, Y.; Li, B.; Zhang, C.; Zhang, C.; Wei, Y.; and Sun, J. 2021. End-to-End Human Object Interaction Detection with HOI Transformer. In *CVPR*.