# GEN-VLKT: Simplify Association and Enhance Interaction Understanding for HOI Detection

Yue Liao[1,3*]   Aixi Zhang[2*]   Miao Lu[2]   Yongliang Wang[2]   Xiaobo Li[2]   Si Liu[1,3†]

[1]Institute of Artificial Intelligence, Beihang University    [2]Alibaba Group

[3]Hangzhou Innovation Institute, Beihang University

## Abstract

*The task of Human-Object Interaction (HOI) detection could be divided into two core problems, i.e., human-object association and interaction understanding. In this paper, we reveal and address the disadvantages of the conventional query-driven HOI detectors from the two aspects. For the association, previous two-branch methods suffer from complex and costly post-matching, while single-branch methods ignore the features distinction in different tasks. We propose Guided-Embedding Network (GEN) to attain a two-branch pipeline without post-matching. In GEN, we design an instance decoder to detect humans and objects with two independent query sets and a position Guided Embedding (p-GE) to mark the human and object in the same position as a pair. Besides, we design an interaction decoder to classify interactions, where the interaction queries are made of instance Guided Embeddings (i-GE) generated from the outputs of each instance decoder layer. For the interaction understanding, previous methods suffer from long-tailed distribution and zero-shot discovery. This paper proposes Visual-Linguistic Knowledge Transfer (VLKT) training strategy to enhance interaction understanding by transferring knowledge from a visual-linguistic pre-trained model CLIP. In specific, we extract text embeddings for all labels with CLIP to initialize the classifier and adopt a mimic loss to minimize the visual feature distance between GEN and CLIP. As a result, GEN-VLKT outperforms the state of the art by large margins on multiple datasets, e.g., +5.05 mAP on HICO-Det. The source codes are available at* https://github.com/YueLiao/gen-vlkt.

## 1. Introduction

Human-Object Interaction (HOI) detection is a significant task to make a machine understand human activities in a static image at a fine-grained level. In this task, human activities are represented as a series of HOI triplets <Human,
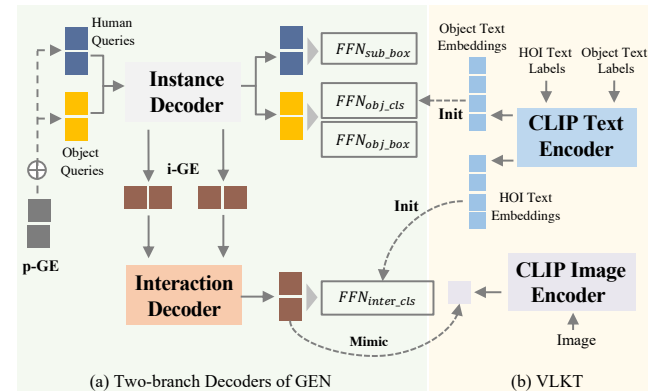


Figure 1. **Our GEN-VLKT pipeline.** We propose GEN, a query-based HOI detector with two-branch decoders, where we design a guided embedding association mechanism to replace the traditional post-matching process for simplifying the association. Moreover, we devise a training strategy VLKT, where we transfer knowledge from the large-scale visual-linguistic pre-trained model CLIP to enhance interaction understanding.

Object, Verb>, so an HOI detector is required to localize human and object pairs and recognize their interactions. The core problems of HOI detection are to explore how to *associate the interactive human and object pairs* and *understand their interactions*. Thus, we consider improving the HOI detector from the two aspects and design a unified and superior HOI detection framework. We first revisit the efforts conducted by traditional methods in such two aspects.

For the *association problem*, it can be mainly divided into two paradigms, *i.e.*, bottom-up and top-down. Bottom-up methods [6, 7, 21] detect humans and objects first and then associate humans and objects through a classifier or a graph model. Top-down methods usually design an anchor to denote the interaction, *e.g.*, interaction point [23] and queries [4, 31, 46], and then find the corresponding human and object through pre-defined associative rules. Benefiting from the development of visual transformer, query-based methods are leading the performance of HOI detection, which are mainly two streams, *i.e.*, two-

---
*Equal contribution

†Corresponding author (liusi@buaa.edu.cn)

branch prediction-then-matching manner [4] and single-branch directly-detection manner [31, 46]. The two-branch manner predicts interaction then matches with human and object, struggling with designing effective matching rules and complicated post-processing. The single-branch manner proposes to detect the human, object and the corresponding interaction based on a single query with multiple heads in an end-to-end manner. However, we argue that the three tasks, *i.e.*, human detection, object detection and interaction understanding, exist significant differences in feature representation, where human and object detection mainly focus on the features in their corresponding regions, while interaction understanding attends human posture or context.

To improve this, as shown in Figure 1a, we propose to keep the two-branch architecture while removing the complicated post-matching. To this end, we propose Guided Embedding Network (GEN), where we adopt an architecture of a visual encoder followed by two-branch decoders, *i.e.*, instance decoder and interaction decoder, and design a guided embedding mechanism to guide the association beforehand. The two branches are both with a query-based transformer decoder architecture. For the instance decoder, we design two independent query sets for human and object detection. Further, we develop a position Guided Embedding (p-GE) to distinguish different human-object pairs by assigning the human query and object query at the same position as a pair. For the interaction decoder, we devise an instance Guided Embedding (i-GE), where we generate each interaction query guided by specific human and object queries to predict its HOIs. Hence, GEN can allow different features for different tasks and guide the association during network forward while without post-matching.

For the *interaction understanding problem*, most conventional methods directly apply a multi-label classifier fitted from the dataset to recognize the HOIs. However, such paradigms suffer from the long-tailed distribution and zero-shot discovery due to the complicated human activities with various interactive objects in realistic scenes. Though recent methods propose to alleviate such problems with data-augmentation [14] or carefully designed loss [44], the performance gain and extension ability are restricted to the limited training scale due to the expensive HOI annotation. We might as well set our sights on image-text data, which can be easily obtained from the internet, while HOI triplets can be naturally converted into text descriptions. Thanks to the development of visual-linguistic pre-trained models [26, 29, 40], especially, CLIP [29] establishes a strong visual-linguistic model trained on about 400 million image-text pairs and shows its powerful generalization ability on about 30 tasks. Thus, CLIP can cover most HOI scenes in real life and bring a new idea to understand HOIs.

To improve this, as shown in Figure 1b, we design a Visual-Linguistic Knowledge Transfer (VLKT) training strategy to transfer the knowledge from CLIP to the HOI detector to enhance interaction understanding without additional computation cost. We consider two main problems in our VLKT. On the one hand, we design a text-driven classifier for prior knowledge integration and zero-shot HOI discovery. In detail, we first covert each HOI triplet label into a phrase description, then extract their text embeddings based on the text encoder of CLIP. Finally, we apply the text embeddings of all HOI labels to initialize the weight of the classifier. In this manner, we can easily extend a novel HOI category only by adding its text embedding into the matrix. Meanwhile, we also adopt the CLIP-initialized object classifier for novel object extension. On the other hand, for text-driven classifier and visual feature alignment, we present a knowledge distillation method to guide the visual features of HOI detection to mimic the CLIP features. Therefore, based on VLKT, the model can well capture information from CLIP and easily extend to novel HOI categories without extra cost during inference.

Finally, we propose a novel unified HOI detection framework GEN-VLKT based on the above two designs. We have verified the effectiveness of our GEN-VLKT on two representative HOI detection benchmarks, *i.e.*, HICO-Det [28] and V-COCO [9]. Our GEN-VLKT has significantly improved the existing methods on both two benchmarks and the zero-shot settings of the HICO-Det dataset. Specifically, our GEN-VLKT has achieved a $5.05$ mAP gain on HICO-Det and a $5.28$ AP promotion on V-COCO compared with the previous state-of-the-art method QPIC [31]. It also promotes performance impressively by a $108.12\%$ relative mAP gain for unseen object zero-shot setting compared to the previous state-of-the-art method ATL [14].

## 2. Related Works

**HOI detection.** Conventional HOI detectors are mainly divided into two folds, bottom-up and top-down. The bottom-up pipelines [3, 6, 7, 10, 18, 19, 21, 24, 34, 37, 45] first detect all humans and objects and then associate the human-object pairs and infer their HOI types through an additional classifier. These methods are usually organized as a two-stage paradigm and worked on improving the second stage. Recently, some graph-based methods [28, 33, 35, 39, 45] have achieved satisfactory performance. However, bottom-up methods suffer from expensive computation consume due to its serial architecture for handling a large number of human-object pairs. To alleviate this issue, top-down methods become popular in recent works [4, 16, 17, 23, 31, 36, 46]. Top-down methods mainly design an additional anchor to associate humans and objects, and predict their interactions. The interaction anchor is from the early interaction point [23, 36] and union box [16] to recent interaction query [4, 17, 22, 31, 46] with the development of visual transformers. Recently, CDN [41] proposed a one-stage method
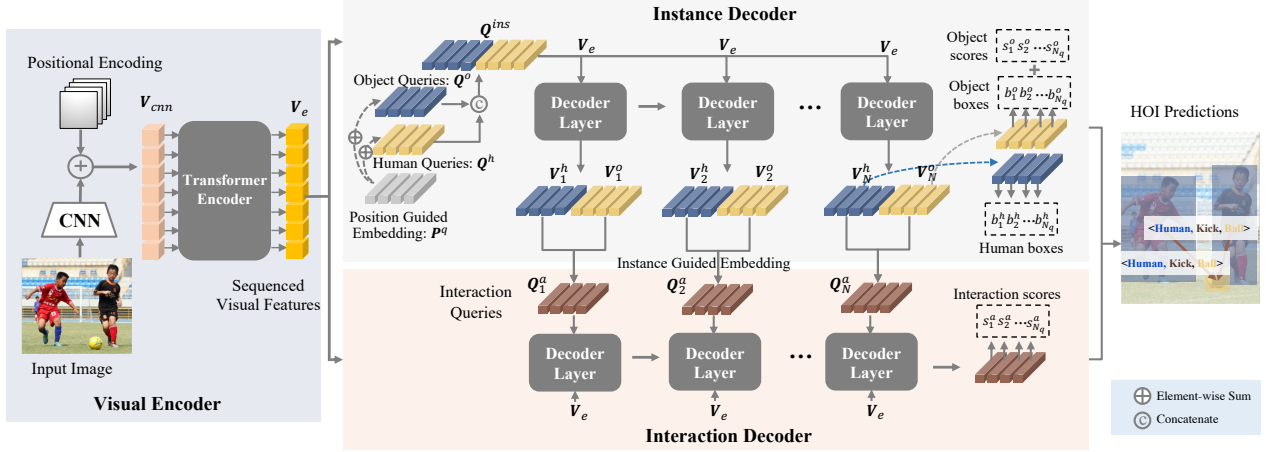
Figure 2. **The framework of our GEN**. The GEN is organized as a visual encoder equipped with two-branch decoders architecture. Given an image, the visual encoder is first applied to extract the visual features. Then, two branches, *i.e.*, instance decoder and interaction decoder, are used to localize human-object pairs and classify HOI triplets based on learnable queries, respectively. Besides, we design a position Guided Embedding (p-GE) to associate the interactive human and object and an instance Guided Embedding (i-GE) to make the interaction query predict the corresponding HOI categories under the guidance of specific human and object queries.

with a cascade decoder to mine the benefits of the above two folds. Our GEN is different from CDN in the three aspects. 1) Organization of the decoder: GEN is with a two-branch pipeline and the instance and interaction decoder forward together, while CDN disentangles the HOI detection into two serial decoders. 2) Instance query design: GEN adopts two isolated human and object queries with positional embedding, while CDN entangles human and object into a unified instance query. 3) motivation: GEN aims to replace the complex post-process with a guided learning manner, while CDN aims to mine the benefits of the two folds.

**Zero-shot HOI detection.** Zero-shot HOI detection [30] tends to detect unseen HOIs in the training data. Many methods [1, 11, 13–15, 25, 27, 30, 32, 38] are investigated to handle zero-shot HOI detection. In detail, [1,11,30,32] factorized the human and object features by disentangled reasoning on verbs and objects and then produced novel HOI triplets during inference. VCL [13] composed novel HOI samples by combining decomposed object and verb features with pair-wise images and within images. FCL [15] presented an object fabricator to generate fake object representations for rare and unseen HOIs. ATL [14] explored object affordances from additional object images to discover novel HOI categories. ConsNet [25] explicitly encoded the relations among objects, actions and interactions into an undirected graph to propagate knowledge among HOI categories as well as their constituents. The visual-linguistic models [27, 38] transferred the seen visual phrase embeddings with prior language knowledge to unseen HOIs.

## 3. Methods

In this section, we aim to explore the solutions for the two problems of HOI detection, *i.e.*, association and in-

teraction understanding. We first present a detailed introduction of our one-stage two-branch HOI detector with a simple association mechanism, namely Guided Embedding Network (GEN), in Sec 3.1. We then introduce a Visual-Linguistic Knowledge Transfer (VLKT) training strategy with the large-scale visual-linguistic pre-trained model CLIP to enhance interaction understanding in Sec 3.2. Finally, we show the training and inference pipelines.

### 3.1. Guided Embedding Network

In this subsection, we introduce the architecture of our Guided Embedding Network (GEN). As shown in Figure 2, the GEN is organized as an encoder followed by two-branch decoders architecture. We first adopt a CNN equipped with a transformer encoder architecture as the visual encoder to extract sequenced visual features $V_e$. Then, we apply two-branch decoders, *i.e.*, instance decoder and interaction decoder, to detect HOI triplets. In the instance decoder, based on $V_e$, we detect humans and objects through the human query set $Q^h$ and the object query set $Q^o$ individually. Additionally, we design a position Guided Embedding (p-GE) $P^q$ to assign the human and object queries at the same position as a pair. In the interaction decoder, we first dynamically generate the interaction queries $Q_i^a$ for each interaction decoder layer by computing the mean of the outputs of human and object queries in the corresponding instance decoder layer. Therefore, the interaction decoder can predict the corresponding HOI categories under the guidance of human and object queries. Finally, the HOI prediction results are generated by the output of decoders.

**Visual Encoder.** We follow the query-based transformer detectors [2, 31, 46] to adopt a CNN-transformer combined architecture for the visual encoder. Taking an image $I$ as
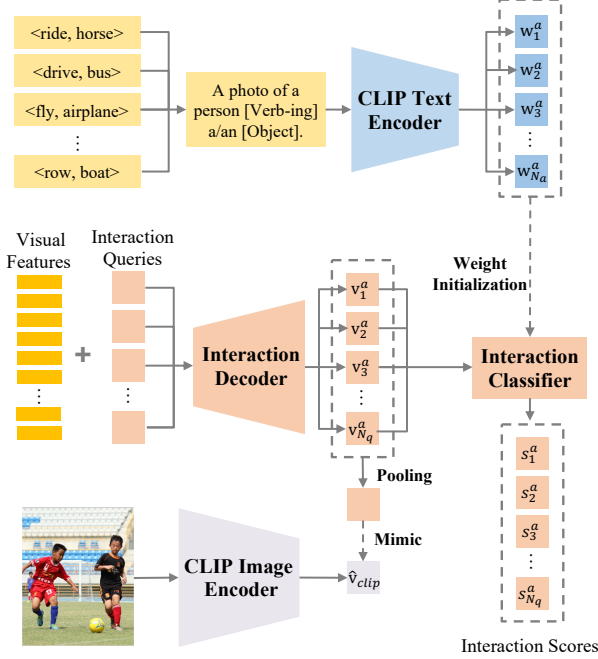
Figure 3. **VLKT for the interaction decoder**. We first covert each HOI label into a phrase description and extract its text embedding based on CLIP text encoder, then apply the text embeddings of all HOI labels to initialize the classifier. Finally, we adopt the CLIP image encoder to extract visual features to guide the interaction visual feature learning. The dotted arrow denotes no gradient.

input, a CNN is first utilized to extract low-resolution visual features $\boldsymbol{V}_{cnn} \in \mathbb{R}^{H' \times W' \times C'}$. Then, we reduce the channels of visual features to $C_e$ and flatten the size of the features to $(H' \times W') \times C_e$. Finally, we feed the reduced features adding a cosine positional embedding into a transformer encoder, and extract sequenced visual features $\boldsymbol{V}_e \in \mathbb{R}^{(H' \times W') \times C_e}$ for the following tasks.

**Decoders.** The decoders in the two branches share the same architecture, where we follow the transformer-based detectors [2, 4] to adopt the query-based transformer decoder framework. First, we feed a set of learnable queries $\boldsymbol{Q} \in \mathbb{R}^{N_q \times C_q}$, the output of last layer, the visual features $V_e$ and the positional embedding to $N$ transformer decoder layers, and output the updated queries after the self-attention and co-attention operations. Then with separate FFN heads, the queries are transformed to embeddings for its dedicated task, *i.e.*, instance and interaction representations by the first and second decoder branch, respectively.

For the instance decoder, we first initialize two sets of queries to detect human and object [5] where we denote human and object query sets as $\boldsymbol{Q}^h \in \mathbb{R}^{N_q \times C_q}$ and $\boldsymbol{Q}^o \in \mathbb{R}^{N_q \times C_q}$ separately. Then, we design an additional learnable position Guided Embedding (p-GE) $\boldsymbol{P}^q \in \mathbb{R}^{N_q \times C_q}$ for the two query sets to assign the human query and object query at the same position as a pair, where we add the $\boldsymbol{P}^q$ to

$\boldsymbol{Q}^h$ and $\boldsymbol{Q}^o$, respectively. Finally, we generate the query set for the instance decoder by concatenating the added queries:

$$\boldsymbol{Q}^{ins} = [\boldsymbol{Q}^h + \boldsymbol{P}^q, \boldsymbol{Q}^o + \boldsymbol{P}^q], \tag{1}$$

where $\boldsymbol{Q}^{ins} \in \mathbb{R}^{2N_q \times C_q}$. We feed $\boldsymbol{Q}^{ins}$ forward the instance decoder to predict the human-object bounding-box pairs $(\boldsymbol{b}_i^h, \boldsymbol{b}_i^o, \boldsymbol{s}_i^o)$, where $\boldsymbol{b}_i^h \in \boldsymbol{B}^h$, $\boldsymbol{b}_i^o \in \boldsymbol{B}^o$ and $\boldsymbol{s}_i^o \in \boldsymbol{S}^o$ denote human bounding-box, object bounding-box and object category scores. And we extract the middle features decoded by each decoder layer as $\mathbf{V}^{ins} = [\mathbf{V}^h, \mathbf{V}^o]$ for the following interaction decoder, where $\mathbf{V}^{ins} \in \mathbb{R}^{N \times 2N_q \times C_q}$.

The goal of the interaction decoder is to predict the HOI categories for the corresponding human-object pair. Therefore, this branch is required to associate interaction query with the human-object query pairs and classify interaction. Here, we introduce an instance Guided Embedding (i-GE) method for the association, and the new interaction classification manner will be introduced in the next subsection. Instead of conventional learnable embedding with random initialization, we dynamically generate i-GE as the interaction queries to guide the interaction query to match the human and object queries. In this manner, we generate i-GE under the guidance of the middle visual features $[\mathbf{V}^h, \mathbf{V}^o]$. Specifically, for the input of $k$-th layer interaction decoder, the interaction queries $\boldsymbol{Q}_k^a$ is computed by the outputs of $k$-th layer instances decoder:

$$\boldsymbol{Q}_k^a = (\boldsymbol{V}_k^h + \boldsymbol{V}_k^o)/2. \tag{2}$$

In this way, the $k$-th layer interaction decoder takes the interaction queries $\boldsymbol{Q}_k^a$ as input, and return the immediate decoded features $\mathbf{V}_k^a$ and HOI categories.

### 3.2. Visual-Linguistic Knowledge Transfer

In this subsection, we detailedly introduce the training pipeline of the instance decoder and the interaction decoder transferring knowledge from the large-scale visual-linguistic pre-trained model CLIP [29], namely Visual-Linguistic Knowledge Transfer (VLKT). In VLKT, we first introduce the process of adopting the CLIP text embeddings to classify interactions and objects. We then introduce how to transfer the visual knowledge from CLIP image embedding to the interaction decoder. We present the pipeline of interaction decoder training with VLKT in Figure 3.

**Text Embedding for Classifier Initialization.** To generate the CLIP text embedding, we first convert HOI triplet labels and object labels into text descriptions. For example, given an HOI triplet <Human, Object, Verb>, we generate the corresponding description following such format 'A photo of a person [Verb-ing] a/an [Object]'. In addition, the 'no-interaction' type is represented as 'A photo of a person and a/an [Object]'. As for an object label, we transform it into the phrase 'A photo of a/an [Object]'. Then, we generate the text embedding for each HOI and object text label

through the pre-trained CLIP text-encoder offline. Finally, the text embedding set of HOI labels $\boldsymbol{E}^a \in \mathbb{R}^{c_t \times N_a}$ and object labels $\boldsymbol{E}^o \in \mathbb{R}^{c_t \times N_o}$ are obtained, where $N_a$ and $N_o$ denote the number of HOI triplet categories and object categories, respectively, and $c_t$ represents the dimension of text embedding from CLIP text encoder.

After obtaining the text embeddings, we aim to classify the interaction and object under the guidance of the prior knowledge from such text embeddings. The intuitive idea is to adopt such embeddings to initialize the weight of the learnable classifier and fine-tune the classifier with a small learning rate to fit a specific dataset. In this way, each output query feature is computed cosine similarity with all fine-tuned text embeddings and returned a similarity score vector during the classification process. Specifically, we denote the interaction classifier and object classifier as $\mathcal{C}^a \leftarrow \boldsymbol{E}^a$ and $\mathcal{C}^o \leftarrow \boldsymbol{E}^o$, respectively. Taking interaction classifier $\mathcal{C}^a = [\mathbf{w}_1^a, \mathbf{w}_2^a, ..., \mathbf{w}_{N_a}^a]$ as an example, given an output interaction query $\mathbf{v}_i^a$, we compute the similarity score by:

$$\mathbf{s}_i^a = \theta \left[ \text{sim}\left(\mathbf{v}_i^a, \mathbf{w}_1^a\right), \text{sim}\left(\mathbf{v}_i^a, \mathbf{w}_2^a\right), \cdots, \text{sim}\left(\mathbf{v}_i^a, \mathbf{w}_{N_a}^a\right) \right] \quad (3)$$

where $\text{sim}$ denotes the cosine similarity operation, for example $\text{sim}\left(\mathbf{v}_i^a, \mathbf{w}_1^a\right) = (\mathbf{v}_i^a \cdot \mathbf{w}_1^a)/(\|\mathbf{v}_i^a\|\|\mathbf{w}_1^a\|)$, and $\theta$ is a logit scale factor following CLIP [29]. The object classification scores can be got in the same way. Otherwise, we follow [31] to apply the focal loss and cross-entropy loss to train the interaction and object classifier, respectively.

**Visual Embedding Mimic.** CLIP is trained on image-text pair data, and it aligns visual embedding and text embedding into a unified space. We design a Visual Embedding Mimicking mechanism to pull the interaction feature into such unified space by pulling the distance between the interaction feature and CLIP visual embedding. Here, CLIP serves as the teacher, and the interaction decoder plays the student's role. We design the knowledge distillation strategy from the global image level, because CLIP image encoder is built upon a whole image. We first feed the resized and cropped image into the pre-trained CLIP visual encoder and extract the visual embedding $\hat{\mathbf{v}}_{clip}$ for the teacher supervision. The global student visual embedding is generated by conducting an average pooling among all output interaction query features. $\mathcal{L}_1$ loss is utilized to pull the distance between the student and the teacher. We formulate the global knowledge distillation as:

$$\mathcal{L}_{glo} = |\hat{\mathbf{v}}_{clip} - \frac{1}{N_q}\sum_{i=1}^{N_q}\mathbf{v}_i^a|, \quad (4)$$

where $N_q$ denotes the number of queries.

### 3.3. Training and Inference

Here, we show the processes of training and inference.

**Training.** During the training stage, we follow the query-based methods [2,31,46] to assign a bipartite matching pre-

diction with each ground-truth using the Hungarian algorithm. The matching process combines the predictions from the FFN heads of the two-branch decoders since the queries of human, object and interaction are one-to-one corresponding. The matching cost for the matching process and the targeting cost for the training back-propagation share the same strategy, where is composed by the box regression loss $\mathcal{L}_b$, the intersection-over-union loss $\mathcal{L}_u$ and the classification loss $\mathcal{L}_c$. The cost is formulated as:

$$\mathcal{L}_{cost} = \lambda_b \sum_{i \in (h,o)} \mathcal{L}_b^i + \lambda_u \sum_{j \in (h,o)} \mathcal{L}_u^j + \sum_{k \in (o,a)} \lambda_c^k \mathcal{L}_c^k, \quad (5)$$

where $\lambda_b$, $\lambda_u$ and $\lambda_c^k$ are the hyper-parameters for adjusting the weights of each loss. Then, considering the mimic loss, the final training loss is given as:

$$\mathcal{L} = \mathcal{L}_{cost} + \lambda_{mimic}\mathcal{L}_{glo}, \quad (6)$$

where $\lambda_{mimic}$ is the hyper-parameter weight for distilling the image embeddings. Additionally, we apply an intermediate supervision for the output of each decoder layer.

**Inference.** The visual embedding mimic only contributes the training stage, and we remove it during inference. For each human-object bounding-box pair $(b_i^h, b_i^o)$ with the object score $s_i^o$ from instance decoder branch, the interaction score is predicted as $s_i^a$ from the interaction decoder. Then, we extend $s_i^o$ from $N_o$-dim to $N_a$-dim, where the score for specific object category will be copy-paste several times for its all corresponding HOI categories. The HOI triplet score is given as $s_i^a + s_i^o(N_a)s_i^o(N_a)$ instead of $s_i^a s_i^o(N_o)$ for balancing the weights of interaction score and object score. The HOI triplets with top $K$ confidence scores are preserved as the final predictions.

## 4. Experiments

In this section, we demonstrate the effectiveness of our designed GEN-VLKT with comprehensive experiments.

### 4.1. Experimental Setting

**Datasets.** We evaluate our model on two public benchmarks, HICO-Det [3] and V-COCO [9]. HICO-Det has $47,776$ images ($38,118$ for training and $9,658$ for testing). It contains $600$ classes of HOI triplets constructed by $80$ object categories and $117$ action categories. V-COCO is a subset of COCO dataset and has $10,396$ images ($5,400$ for training and $4,964$ for testing). It has $29$ action categories which includes $4$ body motions without interaction to any objects. It has the same $80$ object categories. Its actions and objects form $263$ classes of HOI triplets.

**Data Structure for Zero-Shot.** For zero-shot HOI detection, we conduct experiments on HICO-Det following the setting in [1]: 1) Unseen Composition (UC) and 2) Unseen Object (UO). Specifically, the UC setting indicates the training data contains all categories of object and verb but misses

| Method | Detector | Backbone | Anchor | Default | | | Know Object | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Full | Rare | Non-Rare | Full | Rare | Non-Rare |
| Bottom-up Methods: | | | | | | | | | |
| InteractNet [8] | COCO | ResNet-50-FPN | ✗ | 9.94 | 7.16 | 10.77 | - | - | - |
| GPNN [28] | COCO | Res-DCN-152 | ✗ | 13.11 | 9.34 | 14.23 | - | - | - |
| iCAN [7] | COCO | ResNet-50 | ✗ | 14.84 | 10.45 | 16.15 | 16.26 | 11.33 | 17.73 |
| No-Frills [10] | COCO | ResNet-152 | ✗ | 17.18 | 12.17 | 18.68 | - | - | - |
| PMFNet [34] | COCO | ResNet-50-FPN | ✗ | 17.46 | 15.65 | 18.00 | 20.34 | 17.47 | 21.20 |
| DRG [6] | COCO | ResNet-50-FPN | ✗ | 19.26 | 17.74 | 19.71 | 23.40 | 21.75 | 23.89 |
| VCL [13] | COCO | ResNet-50 | ✗ | 19.43 | 16.55 | 20.29 | 22.00 | 19.09 | 22.87 |
| VSGNet [33] | COCO | ResNet-152 | ✗ | 19.80 | 16.05 | 20.91 | - | - | - |
| FCMNet [24] | COCO | ResNet-50 | ✗ | 20.41 | 17.34 | 21.56 | 22.04 | 18.97 | 23.12 |
| ACP [18] | COCO | ResNet-152 | ✗ | 20.59 | 15.92 | 21.98 | - | - | - |
| PD-Net [43] | COCO | ResNet-152-FPN | ✗ | 20.81 | 15.90 | 22.28 | 24.78 | 18.88 | 26.54 |
| SG2HOI [12] | COCO | ResNet-50 | ✗ | 20.93 | 18.24 | 21.78 | 24.83 | 20.52 | 25.32 |
| DJ-RN [19] | COCO | ResNet-50 | ✗ | 21.34 | 18.53 | 22.18 | 23.69 | 20.64 | 24.60 |
| SCG [42] | COCO | ResNet-50-FPN | ✗ | 21.85 | 18.11 | 22.97 | - | - | - |
| IDN [20] | COCO | ResNet-50 | ✗ | 23.36 | 22.47 | 23.63 | 26.43 | 25.01 | 26.85 |
| ATL [14] | HICO-Det | ResNet-50 | ✗ | 23.81 | 17.43 | 25.72 | 27.38 | 22.09 | 28.96 |
| Top-down Methods: | | | | | | | | | |
| UnionDet [16] | COCO | ResNet-50-FPN | $B$ | 17.58 | 11.72 | 19.33 | 19.76 | 14.68 | 21.27 |
| IP-Net [36] | COCO | Hourglass-104 | $P$ | 19.56 | 12.79 | 21.58 | 22.05 | 15.77 | 23.92 |
| PPDM-Hourglass [23] | HICO-Det | Hourglass-104 | $P$ | 21.94 | 13.97 | 24.32 | 24.81 | 17.09 | 27.12 |
| HOI-Trans [46] | HICO-Det | ResNet-50 | $Q$ | 23.46 | 16.91 | 25.41 | 26.15 | 19.24 | 28.22 |
| GG-Net [44] | HICO-Det | Hourglass-104 | $P$ | 23.47 | 16.48 | 25.60 | 27.36 | 20.23 | 29.48 |
| PST [5] | - | ResNet-50 | $Q$ | 23.93 | 14.98 | 26.60 | 26.42 | 17.61 | 29.05 |
| HOTR [17] | HICO-Det | ResNet-50 | $Q$ | 25.10 | 17.34 | 27.42 | - | - | - |
| AS-Net [4] | HICO-Det | ResNet-50 | $Q$ | 28.87 | 24.25 | 30.25 | 31.74 | 27.07 | 33.14 |
| QPIC [31] | HICO-Det | ResNet-50 | $Q$ | 29.07 | 21.85 | 31.23 | 31.68 | 24.14 | 33.93 |
| GEN-VLKT$_s$ | HICO-Det | ResNet-50 | $Q$ | 33.75 | 29.25 | 35.10 | 36.78 | 32.75 | 37.99 |
| GEN-VLKT$_m$ | HICO-Det | ResNet-101 | $Q$ | 34.78 | **31.50** | 35.77 | 38.07 | **34.94** | 39.01 |
| GEN-VLKT$_l$ | HICO-Det | ResNet-101 | $Q$ | **34.95** | 31.18 | **36.08** | **38.22** | 34.36 | **39.37** |

Table 1. **Performance comparison on the HICO-Det test set.** We present an additional tag 'Anchor' to disgust the interaction anchor types for top-down methods, where the '$B$', '$P$' and '$Q$' denote bounding-box, point and query, respectively.

some HOI triplet categories, while the UO setting means the objects in the unseen triplets also do not appear in the training data. We evaluate the 120 unseen, 480 seen, and 600 full categories for the UC setting. Similar to [13], the Rare First UC (RF-UC) selects unseen categories from tail HOIs preferentially, while the Non-rare First UC (NF-UC) prefers the head categories. For the UO setting, we use the unseen HOIs with 12 objects unseen among the total 80 objects and form 100 unseen and 500 seen HOIs. Besides, for a more comprehensive demonstration of our method to investigate the novel HOIs, we propose an Unseen Verb (UV) setting, where we randomly select 20 verbs from all total 117 verbs to form 84 unseen and 516 seen HOIs.

**Evaluation Metric.** We follow the settings in [3] to use the mean Average Precision (mAP) for evaluation. We define a HOI triplet prediction as a true positive if 1) both predicted human and object bounding-boxes have IoU larger than 0.5 *w.r.t.* the GT boxes; and 2) both predicted HOI categories are accurate. For HICO-Det, we evaluate the three different

category sets: all 600 HOI categories (Full), 138 HOI categories with less than 10 training instances (Rare) and the other 462 HOI categories (Non-Rare). For V-COCO, we report the role mAPs for two scenarios: S1 for the 29 action categories including the 4 body motions and S2 for the 25 action categories without the no-object HOI categories.

**Implementation Details.** We implement three versions of GEN-VLKT. The backbone is ResNet-50 for GEN-VLKT$_s$ and ResNet-101 for GEN-VLKT$_m$ and GEN-VLKT$_l$. $N$ for each decoder of the two branches is 3 for GEN-VLKT$_s$ and GEN-VLKT$_m$ and 6 for GEN-VLKT$_l$. The number of HOI categories $N_a$ is 600 for HICO-Det and 263 for V-COCO. We set the number of queries $N_q$ to 64 and the number of channels $C_e$ and $C_q$ to 256. We optimize our network with AdamW with a weight decay of $10^{-4}$. We train the model for 90 epochs with an initial learning rate of $10^{-4}$ decreased by 10 times at the 60th epoch. The training is initialized with the parameters of MS-COCO trained DETR [2]. For regular HOI training, we fine-tune the CLIP

| Method | Anchor | $AP^{S1}_{role}$ | $AP^{S2}_{role}$ |
|---|---|---|---|
| Bottom-up Methods: | | | |
| InteractNet [8] | ✗ | 40.0 | - |
| GPNN [28] | ✗ | 44.0 | - |
| iCAN [7] | ✗ | 45.3 | 52.4 |
| TIN [21] | ✗ | 47.8 | 54.2 |
| VCL [13] | ✗ | 48.3 | - |
| DRG [6] | ✗ | 51.0 | - |
| IP-Net [36] | ✗ | 51.0 | - |
| VSGNet [33] | ✗ | 51.8 | 57.0 |
| PMFNet [34] | ✗ | 52.0 | - |
| PD-Net [43] | ✗ | 52.6 | - |
| FCMNet [24] | ✗ | 53.1 | - |
| ACP [18] | ✗ | 53.23 | - |
| IDN [20] | ✗ | 53.3 | 60.3 |
| Top-down Methods: | | | |
| UnionDet [16] | $B$ | 47.5 | 56.2 |
| HOI-Trans [46] | $Q$ | 52.9 | - |
| AS-Net [4] | $Q$ | 53.9 | - |
| GG-Net [44] | $P$ | 54.7 | - |
| HOTR [17] | $Q$ | 55.2 | 64.4 |
| QPIC [31] | $Q$ | 58.8 | 61.0 |
| GEN-VLKT$_s$ | $Q$ | 62.41 | 64.46 |
| GEN-VLKT$_m$ | $Q$ | 63.28 | 65.58 |
| GEN-VLKT$_l$ | $Q$ | **63.58** | **65.93** |

Table 2. **Performance comparison on the V-COCO**. The '$B$', '$P$' and '$Q$' denote bounding-box, point and query, respectively.

| Method | Type | Unseen | Seen | Full |
|---|---|---|---|---|
| Shen *et al.* [30] | UC | 5.62 | - | 6.26 |
| FG [1] | UC | 10.93 | 12.60 | 12.26 |
| ConsNet [25] | UC | 16.99 | 20.51 | 19.81 |
| VCL [13] | RF-UC | 10.06 | 24.28 | 21.43 |
| ATL [14] | RF-UC | 9.18 | 24.67 | 21.57 |
| FCL [15] | RF-UC | 13.16 | 24.23 | 22.01 |
| baseline | RF-UC | 12.52 | 32.70 | 28.66 |
| GEN-VLKT$_s$ | RF-UC | **21.36** | **32.91** | **30.56** |
| VCL [13] | NF-UC | 16.22 | 18.52 | 18.06 |
| ATL [14] | NF-UC | 18.25 | 18.78 | 18.67 |
| FCL [15] | NF-UC | 18.66 | 19.55 | 19.37 |
| baseline | NF-UC | 18.71 | 22.53 | 21.76 |
| GEN-VLKT$_s$ | NF-UC | **25.05** | **23.38** | **23.71** |
| FCL* [15] | UO | 0.00 | 13.71 | 11.43 |
| ATL* [14] | UO | 5.05 | 14.69 | 13.08 |
| baseline | UO | 2.92 | 28.56 | 23.99 |
| GEN-VLKT$_s$ | UO | **10.51** | **28.92** | **25.63** |
| baseline | UV | 13.52 | 29.25 | 27.04 |
| GEN-VLKT$_s$ | UV | **20.96** | **30.23** | **28.74** |

Table 3. **Performance comparison for Zero-Shot HOI detection**. RF is short for rare first, NF is short for non-rare first, and UC, UO, UV indicate unseen composition, unseen object and unseen verb settings, respectively. The baseline is the model of '$s$' architecture without VLKT. * means only the detected boxes are used without object identity information from the detector.

text embeddings initialized interaction classifier and object classifier with a small learning rate of $10^{-5}$. We implement the zero-shot HOI experiments on HICO-Det. For better novel HOI categories extension, we freeze the CLIP initialized weights for both interaction and object classifiers. We set the output dimension of the interaction classifier to the number of 'seen' categories during training, while we update this output dimension to the 'full' 600 categories during inference. We set the cost weights $\lambda_b$, $\lambda_u$, $\lambda^o_c$ and $\lambda^a_c$ to 2.5, 1, 1 and 1, respectively, following QPIC [31]. We follow the official CLIP data pre-processing for visual embedding mimic to resize and center-crop the real-timely augmented image to 224 and feed the processed image to the CLIP visual encoder. We set the loss weight $\lambda_{mimic}$ to 20. We conduct all the experiments with a batch size of 16 on 8 Tesla V100 GPUs and CUDA10.2.

### 4.2. Effectiveness for Regular HOI Detection

We use the official evaluation code to compute the mAPs for both HICO-Det and V-COCO. Table 1 and Table 2 show the performance comparisons of GEN-VLKT with the recent bottom-up and top-down HOI detection methods.

For HICO-Det as shown in Table 1, GEN-VLKT$_s$ outperforms the all existing bottom-up and top-down methods with a large margin. In specific, compared to the state-of-the-art top-down method QPIC [31], GEN-VLKT$_s$ achieves a relative 16.10% mAP gain with a margin of mAP 4.68.

Especially for the rare categories, GEN-VLKT$_s$ achieves mAP 29.25, which significantly outperforms AS-Net [4] by a margin of mAP 5.00, even outperforming the 'Full' setting of all existing methods. This is ascribable to the guided embedding design of the GEN architecture and the powerful VLKT for the long-tail categories. From an efficiency perspective, GEN-VLKT$_s$ contains in total 6 decoder layers considering the two branches. Thus it has almost the same number of parameters and flops compared to QPIC, and fewer parameters and flops compared to AS-Net [4] with 12 decoder layers in total. And GEN-VLKT$_l$ achieves a new state-of-the-art performance of mAP 34.95.

For V-COCO, as shown in Table 2, GEN-VLKT$_s$ achieves role AP 62.41 on Scenario 1 and role AP 64.46 on Scenario 2, which also outperform the previous state-of-the-art method QPIC-R50 [31] with margins of mAP 3.61 and 3.46, respectively. The promotion is not as significant as that on HICO-Det, since the training samples of V-COCO is insufficient compared to HICO-Det to train such a large number of 263 categories classification.

### 4.3. Effectiveness for Zero-Shot HOI Detection

We conduct all the experiments with the '$s$' model. We train the model without VLKT strategy as the baseline. As shown in Table 3, GEN-VLKT$_s$ outperforms the baseline and previous methods for all the Unseen Composition (UC), Unseen Object (UO) and Unseen Verb (UV) settings.

| Setting | Full | Rare | Non-Rare |
|---|---|---|---|
| Base-triplet | 30.96 | 22.28 | 33.55 |
| Base-verb | 31.88 | 26.24 | 33.57 |
| - p-GE | 31.23 | 25.38 | 32.98 |
| - i-GE | 30.83 | 23.86 | 32.91 |

(a) **Network Architecture Setting**: Training 's' model without VLKT.

| Strategy | Full | Rare | Non-Rare |
|---|---|---|---|
| Base-triplet | 30.96 | 22.28 | 33.55 |
| + *interaction text* | 31.71 | 26.08 | 33.39 |
| + *object text* | 32.09 | 26.68 | 33.71 |
| + *mimic* | **33.75** | **29.25** | **35.10** |

(b) **Training Strategies of VLKT.** : Ablations for the training strategies of VLKT.

| $L_1$ | $L_2$ | Full | Rare | Non-Rare |
|---|---|---|---|---|
| - | - | 32.09 | 26.68 | 33.71 |
| ✓ | - | **33.75** | **29.25** | **35.10** |
| - | ✓ | 32.41 | 26.66 | 34.14 |
| ✓ | ✓ | 33.10 | 29.24 | 34.25 |

(c) **Mimic Loss Setting**: The choice of losses for mimic.

Table 4. **Ablations**. We conduct experiments on HICO-Det dataset based on 's' model. The mAP in default setting is reported.

**Unseen Composition.** For UC, compared with FCL [15], GEN-VLKT$_s$ achieves $38.85\%$ and $22.41\%$ relative mAP gains on full categories for rare first and non-rare first selections, respectively. Specifically, benefiting from the VLKT mechanism, GEN-VLKT$_s$ still significantly promotes the performance for the unseen categories without the feature factorization and composition among images like VCL [13], FCL [15] and ATL [14]. The improvements mainly come from the utilization of CLIP, as indicated by comparing GEN-VLKT$_s$ to the baseline. For example, for the rare first UC setting, GEN-VLKT$_s$ promotes mAP from 13.16 to 21.36 compared to FCL and promotes mAP by a significant margin of mAP 8.84 compared to the baseline.

**Unseen Object.** We further evaluate GEN-VLKT$_s$ with unseen object, which reflects the ability to investigate human interactions with novel objects. For full and unseen categories, GEN-VLKT$_s$ outperforms ATL [14] by relative $95.95\%$ and $108.12\%$ mAP gains, respectively. Again, the comparison to the baseline indicates VLKT promotes the performance for the unseen categories significantly.

**Unseen Verb.** Finally we propose the UV setting to discovery novel categories of actions, and we argue this reflects the specific characteristic of zero-shot HOI detection. We compare GEN-VLKT$_s$ with the baseline and obtain a significant $55.03\%$ relative promotion for unseen categories.

### 4.4. Ablation Study

In this subsection, we conduct a series of experiments to analyse the effectiveness of our proposed modules and strategies. All experiments are conducted in HICO-Det dataset based on the 's' model.

**Network Architecture Setting.** In this part, we aim to prove the superiority of our designed framework. Thus, we implement two base models with regular training strategies without VLKT. Firstly, we follow the previous query-based methods to adopt a verb classifier with 117 categories for GEN, namely 'Base-verb' in Table 4a. It shows that our 'Base-verb' has achieved 31.88 mAP outperforming all existing HOI detectors. Secondly, we replace the verb classifier with an HOI triplet classifier with 600 categories, namely 'Base-triplet'. Due to the serious long-tailed distribution, it drops a bit mAP compared with 'Base-Verb', especially for 'Rare' HOIs. Additionally, we explore the importance of components in our GEN. On the one hand, we remove p-GE and replace the independent human and object queries with a unified query for two tasks in 'Base-verb', and the result has dropped a 0.65 mAP. On the other hand, we remove i-GE and use a learnable embedding with random initialization added by the p-GE for interaction queries in 'Base-verb'. It has lost 1.05 mAP compared with 'Base-verb', but still superior to previous single-branch methods.

**Training Strategies of VLKT.** Here, we verify the effectiveness of the proposed components in VLKT and the results are presented in Table 4b. We take the 'Base-triplet' as the baseline model. We first replace the interaction classifier with the text embedding initialization, and the results are reported as '+*interaction text*', which shows the mAP of 'Rare' HOIs has improved a lot. Thus, the prior knowledge from linguistic prior can alleviate the long-tail distribution. We then further equip this model with the text embedding driven object classifier, causing 0.38 mAP improvement. Finally, we add the mimic loss to transfer visual knowledge from CLIP. All performances have boosted a lot, which proves aligning features is essential.

**Mimic Loss Setting.** We discuss the choice of the mimic loss with two loss types, *i.e.*, $\mathcal{L}_1$ and $\mathcal{L}_2$ losses. As shown in Table 4c, if only equipped with $\mathcal{L}_1$ loss, our model has achieved the best performance and a 1.66 mAP gain, and the performance is much better than only equipped with $\mathcal{L}_2$ loss. If we apply $\mathcal{L}_1$ and $\mathcal{L}_2$ losses at the same time by summation, the performance is also worse than only with $\mathcal{L}_1$. Thus, $\mathcal{L}_1$ is more suitable for the mimic loss.

## 5. Conclusion

We propose a novel framework GEN-VLKT to improve the query-based HOI detectors from two aspects, association and interaction understanding. For association, we design a two-branch framework while removing post-matching by a guided embedding mechanism. For interaction understanding, we design a training strategy VLKT, adopting CLIP to enhance interaction understanding. GEN-VLKT has achieved leading performances on regular and zero-shot settings on HICO-Det and V-COCO datasets.

# References

[1] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. Detecting human-object interactions via functional generalization. In *AAAI*, 2020. 3, 5, 7

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3, 4, 5, 6

[3] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 2, 5, 6

[4] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *CVPR*, 2021. 1, 2, 4, 6, 7

[5] Qi Dong, Zhuowen Tu, Haofu Liao, Yuting Zhang, Vijay Mahadevan, and Stefano Soatto. Visual relationship detection using part-and-sum transformers with composite queries. In *ICCV*, 2021. 4, 6

[6] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *ECCV*, 2020. 1, 2, 6, 7

[7] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *BMVC*, 2018. 1, 2, 6, 7

[8] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018. 6, 7

[9] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 2, 5

[10] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *ICCV*, 2019. 2, 6

[11] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *ICCV*, 2019. 3

[12] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. Exploiting scene graphs for human-object interaction detection. In *ICCV*, 2021. 6

[13] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, 2020. 3, 6, 7, 8

[14] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *CVPR*, 2021. 2, 3, 6, 7, 8

[15] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *CVPR*, 2021. 3, 7, 8

[16] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J. Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *ECCV*, 2020. 2, 6, 7

[17] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. Hotr: End-to-end human-object interaction detection with transformers. In *CVPR*, 2021. 2, 6, 7

[18] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. In *ECCV*, 2020. 2, 6, 7

[19] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *CVPR*, 2020. 2, 6

[20] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu. Hoi analysis: Integrating and decomposing human-object interaction. In *NIPS*, 2020. 6, 7

[21] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yan-Feng Wang, and Cewu Lu. Transferable interactiveness prior for human-object interaction detection. In *CVPR*, 2019. 1, 2, 7

[22] Zhimin Li, Cheng Zou, Yu Zhao, Boxun Li, and Sheng Zhong. Improving human-object interaction detection via phrase learning and label composition. *arXiv preprint arXiv:2112.07383*, 2021. 2

[23] 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 *CVPR*, 2020. 1, 2, 6

[24] Yang Liu, Qingchao Chen, and Andrew Zisserman. Amplifying key cues for human-object-interaction detection. In *ECCV*, 2020. 2, 6, 7

[25] Ye Liu, Junsong Yuan, and Chang Wen Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *MM*, 2020. 3, 7

[26] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NIPS*, 2019. 2

[27] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting unseen visual relations using analogies. In *ICCV*, 2019. 3

[28] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018. 2, 6, 7

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2, 4, 5

[30] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. In *WACV*, 2018. 3, 7

[31] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021. 1, 2, 3, 5, 6, 7

[32] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *CVPR*, 2020. 3

[33] Oytun Ulutan, A S M Iftekhar, and B. S. Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *CVPR*, 2020. 2, 6, 7

[34] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *ICCV*, 2019. 2, 6, 7

[35] Hai Wang, Wei-shi Zheng, and Ling Yingbiao. Contextual heterogeneous graph network for human-object interaction detection. In *ECCV*, 2020. 2

[36] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *CVPR*, 2020. 2, 6, 7

[37] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S. Kankanhalli. Learning to detect human-object interactions with knowledge. In *CVPR*, 2019. 2

[38] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *CVPR*, 2019. 3

[39] Dongming Yang and Yuexian Zou. A graph-based interactive reasoning for human-object interaction detection. In *IJCAI*, 2020. 2

[40] Yi Yang, Yueting Zhuang, and Yunhe Pan. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies. *FITEE*, 2021. 2

[41] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. In *NIPS*, 2021. 2

[42] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *ICCV*, 2021. 6

[43] Xubin Zhong, Changxing Ding, Xian Qu, and Dacheng Tao. Polysemy deciphering network for human-object interaction detection. In *ECCV*, 2020. 6, 7

[44] Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *CVPR*, 2021. 2, 6, 7

[45] Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *ICCV*, 2019. 2

[46] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, and Jian Sun. End-to-end human object interaction detection with hoi transformer. In *CVPR*, 2021. 1, 2, 3, 5, 6, 7