Distillation Using Oracle Queries for Transformer-based Human-Object Interaction Detection

Xian Qu[1] Changxing Ding[1,2*] Xingao Li[1] Xubin Zhong[1] Dacheng Tao[3]
[1] South China University of Technology [2] Pazhou Lab, Guangzhou [3] The University of Sydney
eequxian.scut@mail.scut.edu.cn, chxding@scut.edu.cn
{eexingao,eexubin}@mail.scut.edu.cn, dacheng.tao@gmail.com

Abstract

Transformer-based methods have achieved great success in the field of human-object interaction (HOI) detection. However, these models tend to adopt semantically ambiguous queries, which lowers the transformer's representation learning power. Moreover, there are a very limited number of labeled human-object pairs for most images in existing datasets, which constrains the transformer's set prediction power. To handle the first problem, we propose an efficient knowledge distillation model, named Distillation using Oracle Queries (DOQ), which shares parameters between teacher and student networks. The teacher network adopts oracle queries that are semantically clear and generates high-quality decoder embeddings. By mimicking both the attention maps and decoder embeddings of the teacher network, the representation learning power of the student network is significantly promoted. To address the second problem, we introduce an efficient data augmentation method, named Context-Consistent Stitching (CCS), which generates complicated images online. Each new image is obtained by stitching labeled human-object pairs cropped from multiple training images. By selecting source images with similar context, the new synthesized image is made visually realistic. Our methods significantly promote both the accuracy and training efficiency of transformer-based HOI detection models. Experimental results show that our proposed approach consistently outperforms state-of-the-art methods on three benchmarks: HICO-DET, HOI-A, and V-COCO. Code is available at https://github.com/SherlockHolmes221/DOQ.

1. Introduction

Human-Object Interaction (HOI) detection aims to identify a set of meaningful ⟨*human, interaction, object*⟩ triplets in an image. HOI is fundamental for scene and action understanding, with applications including action prediction [1, 2], scene graph generation [3, 4], and visual question answering [5, 6]. It is also a very challenging task. One reason for this degree of difficulty is that the task in question is composite: for each triplet, it is necessary to determine the location of the human and object instances involved, the object category, and the interactions between the human-object pair. Another major reason lies in HOI's reliance on reasoning, which indicates that visual context is important in determining the interaction categories [23].

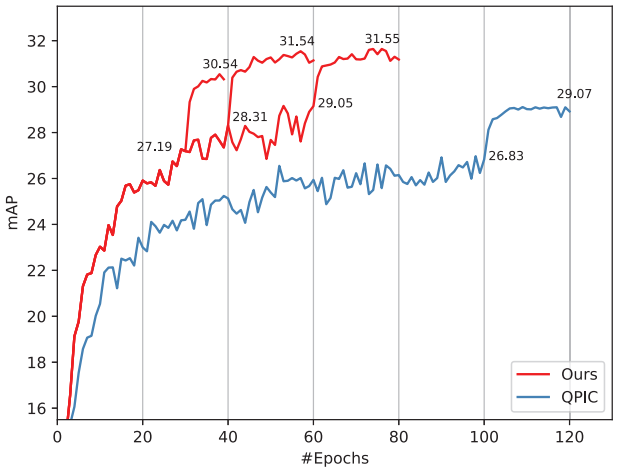

Figure 1. The mAP and convergence curves for QPIC [23] and our model on HICO-DET [48]. Our model achieves better mAP accuracy with a considerably faster convergence rate.

In view of its strong ability to leverage contextual cues, recent methods have applied transformer to HOI detection. For example, QPIC [23] and HOI-Trans [25] defined a set of learnable HOI queries, each of which denotes a possible human-object pair in an image. AS-Net [24] and HOTR [26] defined a set of learnable object and interaction queries to infer object and interaction categories, respectively. Their common ground lies in their utilization of a cross-attention mechanism to mine image-wide contextual information in order to improve HOI detection.

However, the representation learning ability and set prediction power of existing transformer-based HOI detection methods may be underexplored. The first problem lies

*Corresponding author.

in the semantic ambiguity of HOI queries. Each query contains only the rough location of one possible human-object pair [11,25]; thus, the cross-attention operation in the transformer decoder cannot produce precise attention maps, which are essential in acquiring cues from discriminative regions. The representation learning ability is therefore constrained and the convergence rate accordingly slows. The second problem is caused by the limited number of labeled human-object pairs in most training images of existing datasets. Therefore, the transformer may exhibit performance degradation for complicated images with many human-object pairs.

Herein, we propose an efficient knowledge distillation model, named Distillation using Oracle Queries (DOQ), to overcome the first problem. This model adopts one existing transformer-based model, e.g., QPIC [23], as the student network. It shares transformer parameters between teacher and student networks. The differences between the two networks lie in their HOI queries and initial decoder embeddings. For the student network, the HOI queries and initial decoder embeddings are defined as a set of learnable embeddings and zero vectors, respectively. For the teacher network, we construct a set of oracle HOI queries using the ground-truth positions of labeled human-object pairs. We further generate the initial decoder embedding according to the word embedding of the ground-truth object category involved in each labeled human-object pair. In this way, the teacher network acquires both precise semantic and position information for each labeled human-object pair, which enables it to produce high-quality representations and precise attention maps. The representation learning power of the student network is thus significantly promoted due to its mimicking both the attention maps and representations of the teacher network. In terms of inference stage, moreover, the teacher network is abandoned and therefore introduces no additional computational cost.

To address the second problem, we introduce an efficient data augmentation method, named Context-Consistent Stitching (CCS), which creates images online that contain more human-object pairs. In more detail, each new image is obtained by stitching together labeled human-object pairs that are cropped from multiple training images with similar visual context. There are two key advantages of this strategy. First, each synthesized image contains more human-object pairs and does not require manual labeling. Second, by cropping patches from images with similar scenes, the newly created new image is made visually realistic, which is proven to be essential in our experiments. Finally, through the inclusion of the synthesized images, the set prediction power of the transformer is sufficiently optimized.

To the best of our knowledge, our proposed method is the first approach to explicitly handle the semantic ambiguity problem of queries for transformers in HOI detection. We creatively introduce knowledge distillation to address this problem. We demonstrate the effectiveness of our proposed approaches through comprehensive experiments on three HOI detection benchmarks: HICO-DET [48], HOI-A [29], and V-COCO [49], and find that our method consistently achieves state-of-the-art performance. Moreover, benefiting from the knowledge distillation based on oracle queries, our approach achieves a significantly faster convergence rate than existing methods, as illustrated in Figure 1.

2. Related Work

Human-Object Interaction Detection. Existing approaches to HOI detection can be grouped into two paradigms, i.e., two-stage strategies and one-stage strategies. Two-stage methods [21, 28, 34, 41, 43, 44, 46] perform object detection before interaction prediction. Most two-stage approaches adopt a generic object detector and focus on improving interaction prediction. Various types of features can be utilized for interaction prediction, including visual features [31, 45], spatial features [32, 33], human poses [35, 42], and language features [35, 39]. However, due to their sequential structure and redundant human-object instance combinations, two-stage methods frequently encounter the problem of low efficiency.

One-stage HOI detection methods usually perform object detection and interaction prediction in parallel. In the absence of explicit object locations, these methods rely on predefined interaction areas for interaction prediction. Depending on the definition of interaction area employed, existing approaches can be classified into (i) point-based methods, (ii) union region-based methods, and (iii) spatial attention-based methods. Point-based methods adopt a single interaction point [29, 30] or a point set [22] as the interaction area, while union region-based methods [40] regard the union box of a human-object pair as the interaction area. Recently, some approaches have predicted a spatial attention map for each human-object pair as the interaction area, which is achieved by employing cross-attention operations in transformer decoder layers. Spatial attention maps can more flexibly leverage image-wide contextual cues. Transformer-based approaches can be further subdivided into three categories: (i) methods that adopt a set of learnable HOI queries, each of which represents a possible human-object pair [23, 25]; (ii) methods that employ two sets of learnable queries for object detection and interaction prediction, respectively [24, 26]; (iii) methods that define three sets of learnable queries, representing the subject, interaction, and object, respectively [19].

We here address two underexplored issues in existing transformer-based HOI detection methods, namely the constrained representation learning ability and constrained set prediction power. By resolving these two problems, both the accuracy and training efficiency of existing transformer-based approaches are significantly promoted.

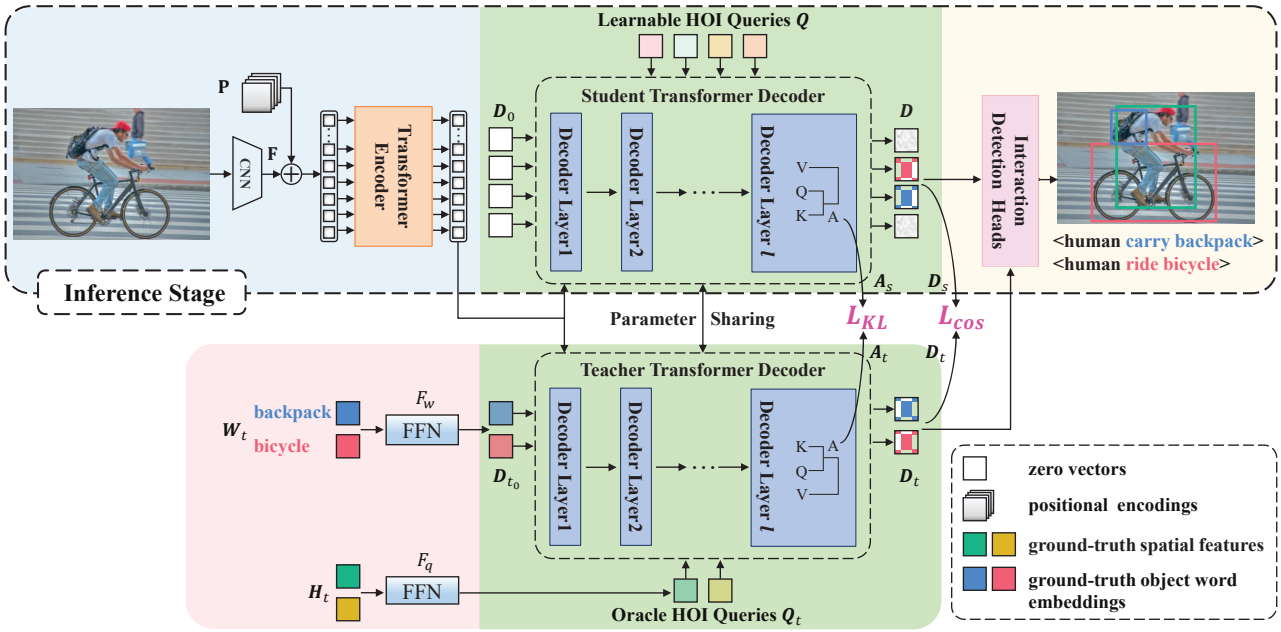

Figure 2. Overview of the proposed approach in the training stage. QPIC [23] is adopted as the student network, which comprises of a CNN backbone, a transformer encoder, a transformer decoder, and interaction detection heads. For the teacher network, we construct a set of oracle HOI queries $\boldsymbol{Q}_t$ according to the ground-truth position of the labeled human-object (HO) pairs. We also generate the initial decoder embedding $\boldsymbol{D}_{t_0}$ according to the word embedding of the ground-truth object category involved in each labeled HO pair. These two networks share parameters. Both the attention maps and the representations of the student network mimic those of the teacher network. In the inference stage, the teacher is abandoned, meaning that no additional computational cost is introduced. Best viewed in color.

Transformer-based Object Detection. The Detection Transformer (DETR) [11] successfully applied transformer to object detection. Many methods that boost the performance of DETR have subsequently emerged: these can be classified into (i) approaches that accelerate training convergence and (ii) approaches that reduce computational complexity. Training convergence can be improved by applying spatial priors to the attention maps of each transformer decoder layer [12, 14]. These priors are usually obtained according to the object location estimated by the former decoder layer. The result can be a map of the same size as the attention maps [12] or a Region of Interest (ROI) [14]. Moreover, Meng et al. [13] disentangled the content and spatial information in queries to generate attention maps, thereby reducing the training difficulty. To handle the problem of high computational complexity, Zhu et al. [10] proposed the deformable attention module, which attends to a small set of sampling locations designated as prominent key elements rather than all pixels in the feature map. Jiang et al. [15] obtained high-resolution attention maps by interpolating available low-resolution ones, thereby considerably reducing the amount of redundant computations required.

Rather than using estimated object locations, we here opt to employ the spatial attention maps produced by oracle queries as spatial priors in the training stage. Our spatial prior is imposed on the entire attention map as supervision, including both objects and context. In the experimentation section, we demonstrate that our proposed approach is more suitable for HOI detection.

3. Methods

Our methods can be applied to many existing transformer-based HOI detection models [23–26, 37]. In this section, we take the representative work QPIC [23] as an example (see illustration in Figure 2). We first revisit its architecture in Section 3.1, then introduce the proposed knowledge distillation model in Section 3.2. Finally, the data augmentation method is described in Section 3.3.

3.1. QPIC Revisited

QPIC builds upon DETR [11] and makes parallel predictions for HOI triplets. It consists of a Convolutional Neural Network (CNN) backbone, a transformer encoder, a transformer decoder, and interaction detection heads. An image is first sent into a CNN backbone to yield visual feature maps $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$; here, C, H, and W denote the number of channels, the height and the width of $\mathbf{F}$, respectively. $\mathbf{F}$ is then augmented with positional encodings $\mathbf{P} \in \mathbb{R}^{C\times H\times W}$ and fed into the transformer encoder, which produces feature maps $\mathbf{E} \in \mathbb{R}^{C\times H\times W}$. The transformer decoder performs cross-attention between a set of learnable HOI queries $\boldsymbol{Q} = \{\mathbf{q}_i | \mathbf{q}_i \in \mathbb{R}^d\}_{i=1}^{N_q}$ and $\mathbf{E}$, where N_q is the number of HOI queries and d denotes their dimension. We represent the operations of the decoder as follows:

$$\boldsymbol{D} = f_{dec}(\boldsymbol{Q}, \boldsymbol{D}_0, \mathbf{E}, \mathbf{P}), \tag{1}$$

where $f_{dec}(\cdot,\cdot,\cdot,)$ is a set of stacked transformer decoder layers. $\boldsymbol{D}_0 = \{\mathbf{d}_{i_0}|\mathbf{d}_{i_0} \in \mathbb{R}^d\}_{i=1}^{N_q}$ denotes the initial decoder embeddings, while $\boldsymbol{D} = \{\mathbf{d}_i|\mathbf{d}_i \in \mathbb{R}^d\}_{i=1}^{N_q}$ denotes the embeddings output by the last decoder layer.

Finally, $\boldsymbol{D}$ is fed into interaction detection heads based on feed-forward networks (FFNs). These four heads are responsible for human bounding box localization, object bounding box localization, object classification, and interaction category prediction, respectively.

3.2. Distillation Using Oracle Queries

The quality of $\boldsymbol{D}$ is vital for reliable HOI detection. However, as revealed in Eq. (1), its discriminative power is affected by $\boldsymbol{Q}$ and $\boldsymbol{D}_0$. Specifically, $\boldsymbol{Q}$ is the same for all images, while $\boldsymbol{D}_0$ is a set of zero vectors, meaning that they are both semantically ambiguous. This problem constrains the representation power of $\boldsymbol{D}$ and slows the convergence of DETR-based models.

To address this issue, we design an efficient knowledge distillation model named Distillation using Oracle Queries (DOQ), the overall structure of which is illustrated in Figure 2. There are two networks involved, namely one teacher network and one student network. The parameters of the CNN backbone, the transformer encoder and decoder, and the interaction detection heads are shared between the two networks. The main differences between them lie in their HOI queries and the initial decoder embeddings. In more detail, the teacher network adopts oracle queries and semantically clear initial decoder embeddings, which enable it to generate precise attention maps and thus discriminative decoder embeddings. We then mimic both the attention maps and decoder embeddings of the teacher network to improve the student's representation learning power. Moreover, as the teacher network only exists during training, there is no additional computational cost in the inference stage.

Oracle HOI Queries. Recent studies [11, 13, 25, 26] show that the HOI or object queries reflect the position information of one possible target. We therefore construct each oracle HOI query according to the ground-truth position of one labeled human-object (HO) pair. With each oracle query, the cross-attention operations in the teacher network accurately attend to one specific HO pair. More specifically, we obtain the set of oracle HOI queries $\boldsymbol{Q}_t$ for one training image as follows:

$$\boldsymbol{Q}_t = tanh(F_q(\boldsymbol{H}_t)), \tag{2}$$

where $\boldsymbol{H}_t = \{\mathbf{h}_{t_i}|\mathbf{h}_{t_i} \in \mathbb{R}^{12}\}_{i=1}^{N_{t_q}}$ and

$$\begin{aligned}\mathbf{h}_{t_i} = [&x_{s_i}, y_{s_i}, w_{s_i}, h_{s_i}, x_{o_i}, y_{o_i}, w_{o_i}, h_{o_i},\\ &x_{s_i} - x_{o_i}, y_{s_i} - y_{o_i}, w_{s_i}h_{s_i}, w_{o_i}h_{o_i}]^T.\end{aligned} \tag{3}$$

N_{t_q} is the number of labeled HO pairs for the image and $\boldsymbol{H}_t$ denotes a set of spatial features. The first eight elements in $\mathbf{h}_{t_i}$ are the center coordinates, width, and height of the human and object bounding boxes of the i-th pair, respectively. $[x_{s_i} - x_{o_i}, y_{s_i} - y_{o_i}]$ represents the relative position [42] between the two boxes, while the last two elements respectively represent the areas of the two boxes. F_q is a two-layer FFN with ReLU that projects $\boldsymbol{H}_t$ to a d-dimensional space. We employ a $tanh$ function for normalization, ensuring the amplitudes of elements in $\boldsymbol{Q}_t$ and $\mathbf{P}$ are consistent.

Initial Decoder Embeddings. The oracle HOI queries contain only box-level position information, which is still coarse for our purposes. In the following, we further enhance the power of the teacher network with improved initial decoder embeddings. More specifically, we generate initial decoder embeddings $\boldsymbol{D}_{t_0} = \{\mathbf{d}_{t_{i_0}}|\mathbf{d}_{t_{i_0}} \in \mathbb{R}^d\}_{i=1}^{N_{t_q}}$ with reference to the word embeddings of the ground-truth object category involved in labeled HO pairs. Formally,

$$\boldsymbol{D}_{t_0} = F_w(\boldsymbol{W}_t), \tag{4}$$

where $\boldsymbol{W}_t = \{\mathbf{w}_{t_i}|\mathbf{w}_{t_i} \in \mathbb{R}^{512}\}_{i=1}^{N_{t_q}}$, and $\mathbf{w}_{t_i}$ denotes the i-th word embedding. F_w denotes another two-layer FFN with ReLU.

Finally, the operations of the transformer decoder in the teacher network can be summarized as follows:

$$\boldsymbol{D}_t = f_{dec}(\boldsymbol{Q}_t, \boldsymbol{D}_{t_0}, \mathbf{E}, \mathbf{P}), \tag{5}$$

where $\boldsymbol{D}_t = \{\mathbf{d}_{t_i}|\mathbf{d}_{t_i} \in \mathbb{R}^d\}_{i=1}^{N_{t_q}}$ denotes the output decoder embeddings of the teacher network. With the help of semantically clear HOI queries and initial decoder embeddings, the teacher network can successfully produce precise attention maps and therefore output high quality decoder embeddings.

Distillation Loss. We align both the output decoder embeddings and the attention maps between the two networks. In more detail, we first establish correspondences between embeddings in $\boldsymbol{D}$ and those in $\boldsymbol{D}_t$. To achieve this goal, we conduct bipartite matching [61] between the predictions of the student network and the ground-truth for each training image according to the strategy outlined in [23]. We then re-arrange the embeddings in $\boldsymbol{D}$ according to the matched ground-truth HO pairs, and denote the set of matched embeddings in $\boldsymbol{D}$ as $\boldsymbol{D}_s$. For the teacher network, as both the HOI queries and the initial decoder embeddings are semantically clear, each embedding in $\boldsymbol{D}_t$ corresponds strictly to a ground-truth HO pair. Finally, we impose the following distillation loss:

$$L_{dis} = \alpha_1 L_{cos} + \alpha_2 L_{KL}, \tag{6}$$

$$L_{cos} = \frac{1}{N_{t_q}} \sum_{i=1}^{N_{t_q}} (1 - \frac{\mathbf{d}_{t_i}{}^T \mathbf{d}_{s_i}}{\|\mathbf{d}_{t_i}\|_2 \|\mathbf{d}_{s_i}\|_2}), \tag{7}$$

$$L_{KL} = \frac{2}{N_{t_q} l} \sum_{j=l/2+1}^{l} \sum_{i=1}^{N_{t_q}} (\boldsymbol{A}_{t_i}^j (ln(\boldsymbol{A}_{t_i}^j) - ln(\boldsymbol{A}_{s_i}^j))), \tag{8}$$

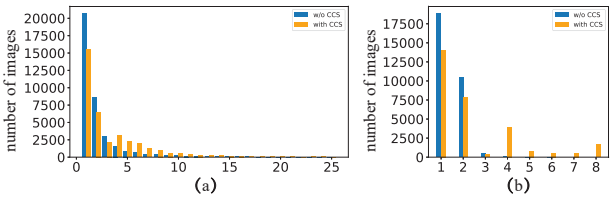

Figure 3. Statistics on the number of labeled HO pairs in existing datasets without CCS (blue color) and with CCS (orange color). (a) HICO-DET [48]. (b) HOI-A [29].

where α_1 and α_2 represent weights, while l is the number of decoder layers, which is set as 6 following [23]. Moreover, $\mathbf{d}_{t_i}$ and $\mathbf{d}_{s_i}$ are the i-th embeddings in $\boldsymbol{D}_t$ and $\boldsymbol{D}_s$, respectively, while $\boldsymbol{A}^j_{t_i}$ and $\boldsymbol{A}^j_{s_i}$ denote the averaged attention maps across the multi-heads in the j-th decoder layer for $\boldsymbol{d}_{t_i}$ and $\mathbf{d}_{s_i}$, respectively. We adopt the KL-divergence loss to align $\boldsymbol{A}^j_{t_i}$ and $\boldsymbol{A}^j_{s_i}$, thereby encouraging the student network to produce attention maps as precise as those of the teacher.

Overall Loss Function. Following [23], we adopt the same loss functions for object detection and interaction prediction. The overall loss function in the training phase is represented as follows:

$$L = L_t + L_s + L_{dis}, \tag{9}$$

where

$$L_t = \lambda_b L_{t_b} + \lambda_u L_{t_u} + \lambda_c L_{t_c} + \lambda_a L_{t_a}, \tag{10}$$

$$L_s = \lambda_b L_{s_b} + \lambda_u L_{s_u} + \lambda_c L_{s_c} + \lambda_a L_{s_a}. \tag{11}$$

L_t and L_s denote the loss functions for the teacher and student networks, respectively. L_{k_b}, L_{k_u}, L_{k_c}, and L_{k_a} ($k \in \{t, s\}$) denote the L1 loss, GIOU loss [59] for bounding box regression, cross-entropy loss for object classification, and focal loss [60] for interaction prediction, respectively; these are realized in the same way as in [23]. Moreover, λ_b, λ_u, λ_c and λ_a are set as 2.5, 1, 1, 1, which are the same values as those in [23].

Discussion. Our model adopts a markedly different approach from most existing knowledge distillation methods [7–9] that train teacher and student models separately. In these works, the teacher model is usually larger in size and therefore achieves better performance, facilitating knowledge distillation to the student model. In comparison, the teacher and student networks in our approach share model parameters; our teacher network achieves higher performance by adopting oracle queries. Moreover, the teacher and student networks in DOQ are trained jointly, which simplifies the training pipeline and thereby significantly reduces training costs.

3.3. Context-Consistent Stitching

We observe that the average number of labeled HO pairs in one image presents a long-tail distribution (shown in Figure 3); that is, most images contain only a small number of labeled HO pairs. There are two main reasons for this. First, the number of HO pairs in most images is indeed small. Second, some HO pairs in complex visual scenes are ignored due to labeling-associated difficulties; this problem constrains the transformer's set prediction power due to insufficient training with images that include many HO pairs. In order to address this problem, we propose an efficient data augmentation method, named Context-Consistent Stitching (CCS), which automatically generates new synthesized images with more labeled HO pairs.

Figure 4. Samples of images synthesized using CCS.

Specifically, in the training phase, we replace each image $\mathbf{I}$ with a synthesized one with probability γ. When an image is chosen for replacement, we first randomly sample three images from its K nearest neighbors in the entire training set. The K neighbors are determined offline according to distances between scene features, which are extracted using an off-the-shelf scene classification model released in [55]. We then crop the union region of one labeled HO pair that is randomly selected from each of the four images, i.e., $\mathbf{I}$ and its three neighbors. If one selected HO pair is found to overlap with other pairs in the same image, we simply extend the union region to include all pairs for which overlap exists. Finally, we stitch these four regions tightly together and resize the stitched image to have a size similar to that of $\mathbf{I}$. We also automatically generate HOI annotations according to the specific locations of each HO pair in new image.

The strategy described above selects images for stitching that are consistent in terms of context, which ensures that the synthesized images are visually realistic. In the experimentation section, we demonstrate that maintaining context consistency is vital for HOI detection. Samples of synthesized images are presented in Figure 4. Moreover, statistics on the number of labeled HO pairs with CCS for existing databases are shown in Figure 3.

It is worth noting that CCS differs from existing copy-and-paste techniques [56–58] used for instance segmentation or object detection, which copy all pixels belonging to selected objects from one image and then paste them into another image. There are two main differences. First, these methods do not consider context consistency across the complete image [56–58]. Second, they are not designed for transformer and do not necessarily change the number of objects in the training images. By contrast, the goal of

CCS is to synthesize images with more labeled HO pairs, by which the set prediction power of transformer-based HOI methods can be improved.

4. Experiments

4.1. Datasets and Metrics

HICO-DET. HICO-DET [48] is a large-scale HOI dataset. It consists of 47,776 images, 38,118 of which are used for training while 9,658 are used for testing. The dataset contains 80 object categories, 117 interaction categories, and 600 HOI categories. Among the 600 HOI categories, there are 138 HOI categories with less than 10 training samples, which are denoted as "rare" categories. There are two evaluation modes: the Default (DT) mode and the Known-Object (KO) mode. HICO-DET uses mean Average Precision (mAP) as its evaluation metric.

HOI-A. The Human-Object Interaction for Application (HOI-A) dataset [29] contains 11 object categories and 10 interaction categories. It comprises of 38,629 images, with 29,842 used for training and 8,787 for testing. The evaluation metrics are the same as for HICO-DET.

V-COCO. The V-COCO dataset [49] is built upon the MS-COCO database [50]. It consists of 10,346 images (5,400 for training and 4,946 for testing), covering 80 object categories and 26 interaction categories. We use the mean average precision of Scenario 1 ($\mathrm{mAP}_{\mathrm{role}}$) [49] for evaluation.

4.2. Implementation Details

We adopt ResNet-50 and ResNet-101 [51] as our backbone model, respectively. We use the AdamW [52] optimizer and conduct experiments with a batch size of 16 on 8 GPUs. The initial learning rate is set to 1e-4 and then multiplied by 0.1 after 60 epochs; the total number of epochs is 80. N_q and d are set as 100 and 256, respectively. We initialize the network with the parameters of DETR [11] trained on MS-COCO database [50]. The word embeddings are extracted by the CLIP model [54] and their dimension is 512. As for the hyper-parameters, α_1 and α_2 in DOQ are set as 1 and 10, respectively; moreover, K and γ in CCS are empirically set as 15 and 0.25, respectively.

4.3. Ablation Study

In the following, we perform ablation studies on both the HICO-DET and HOI-A datasets to demonstrate the effectiveness of DOQ and CCS. Our baseline is QPIC [23]. All experiments are conducted using ResNet-50 as backbone.

Effectiveness of DOQ. As illustrated in Table 1, we first introduce the teacher network, simply sharing transformer parameters between the teacher and student networks without applying the distillation loss. This strategy is referred to as Multi-Task Learning (MTL). It is found that MTL promotes HOI detection performance by 0.56% and 0.76% mAP on HICO-DET and HOI-A, respectively. This is because parameter sharing implicitly aligns the feature space of the two networks. When L_{cos} is introduced, the performance is further improved by 0.65% and 0.55% mAP on HICO-DET and HOI-A, respectively. After L_{KL} is adopted, the performance is further improved by 0.13% and 0.16% mAP on HICO-DET and HOI-A, respectively. The above experiments justify the effectiveness of DOQ.

Table 1. Ablation study on each component of our methods.

Methods	Components				mAP	
	MTL	L_{cos}	L_{KL}	CCS	HICO-DET (DT)	HOI-A
Baseline	-	-	-	-	29.07	74.10
Incremental	✓	-	-	-	29.63	74.86
	✓	-	✓	-	30.13	75.25
	✓	✓	-	-	30.28	75.41
	✓	✓	✓	-	30.41	75.57
	-	-	-	✓	30.76	75.45
Drop-one-out	-	✓	✓	✓	30.82	76.23
	✓	-	✓	✓	31.22	76.57
	✓	✓	-	✓	31.31	76.73
	✓	✓	✓	-	30.41	75.57
Ours	✓	✓	✓	✓	**31.55**	**76.87**

Table 2. Effectiveness of DOQ and CCS on HOTR [26] and CDN [37] in the DT Mode of HICO-DET.

Model	DOQ	CCS	Full	Rare	Non-rare
HOTR	-	-	23.46	16.21	25.62
HOTR	✓	-	25.17	24.15	25.47
HOTR	✓	✓	**25.97**	**26.09**	**25.93**
CDN-S	-	-	31.44	27.39	32.64
CDN-S	✓	-	32.26	27.72	33.62
CDN-S	✓	✓	**33.28**	**29.19**	**34.50**

Effectiveness of CCS. To facilitate clean comparison, we apply CCS to our baseline. We can observe that the performance is promoted by 1.69% and 1.35% mAP on HICO-DET and HOI-A, respectively. We then adopt CCS and DOQ together. It is shown that the performance exceeds the model using DOQ alone by 1.14% and 1.30% on HICO-DET and HOI-A, respectively. We further evaluate the optimal value of γ in CCS; experimental results are provided in the supplementary material.

Drop-one-out Study. We next perform a drop-one-out study in which each proposed component is removed individually. As illustrated by experimental results in Table 1, each proposed component is helpful in promoting HOI detection performance.

Application to Other Transformer-based Methods. DOQ and CCS are plug-and-play and can be readily applied to other transformer-based HOI detection methods (e.g., HOTR [26] and CDN [37]). In the following, we conduct experiments on HICO-DET database to demonstrate the effectiveness of DOQ and CCS on both HOTR and CDN. The detailed network architectures with DOQ are provided in the supplementary material. In the training phase, all settings are kept the same as in their original papers to ensure fair comparison. Results are presented in Table 2. DOQ and CCS are found to improve performance by 1.84% (2.51%), 1.80% (9.88%) and 1.86% (0.31%) mAP in DT mode for the full, rare and non-rare HOI categories respectively on CDN (HOTR). These results show that DOQ and CCS are both portable and flexible.

Table 3. Comparisons with variants of DOQ and CCS in DT Mode of HICO-DET.

(a) **Oracle HOI Queries**

Methods	Full	Rare	Non-Rare
absolute position	30.12	24.16	31.90
Ours	**30.41**	**25.10**	**32.00**

(b) **Initial Decoder Embeddings**

Methods	Full	Rare	Non-Rare
zero vectors	29.70	25.05	31.08
verb-class vectors	30.10	24.23	31.85
object embeddings (Word2Vec)	30.28	23.79	**32.22**
object embeddings (CLIP)	**30.41**	**25.10**	32.00

(c) **Context-Consistent Stitching**

Methods	Full	Rare	Non-Rare
w/o context consistency	29.93	22.66	32.10
w/o stitching diversity	29.97	22.98	32.05
w/o increased HO pairs	29.48	22.10	31.67
Ours	**30.76**	**24.60**	**32.60**

4.4. Comparisons with Variants of DOQ and CCS

In this subsection, we compare the performance of DOQ and CCS with some possible variants. All experiments are conducted in the DT mode of HICO-DET. The results are summarized in Table 3.

Oracle HOI Queries. We generate oracle HOI queries according to $\mathbf{h}_{t_i}$ in Eq. (3). $\mathbf{h}_{t_i}$ is a 12-dimensional vector that includes both the absolute and relative position of the two bounding boxes for a labeled HO pair. Here, we only reserve the first eight elements in $\mathbf{h}_{t_i}$, which means we employ only the absolute position of the two boxes to construct oracle queries. The results listed in Table 3a show that the performance of AP is slightly lower, which informs us that the relative position and the areas of the two bounding boxes are also important for oracle query construction.

Initial Decoder Embeddings. In DOQ, we generate $\boldsymbol{D}_{t_0}$ according to the word embedding of the ground-truth object category involved in each labeled HO pair utilizing CLIP [54]. As illustrated in Table 3b, compared to the use of zero vectors, our design of $\boldsymbol{D}_{t_0}$ boosts the performance by 0.71% mAP. We also attempt to generate $\boldsymbol{D}_{t_0}$ according to the label vector of the verb categories involved in each labeled HO pair. The experimental results in Table 3b show that the word embeddings of objects are more effective at producing $\boldsymbol{D}_{t_0}$. It is further shown that word embeddings extracted using CLIP [54] slightly outperform those obtained using the Word2Vec [53]; this may be because the former was trained to be consistent with the visual features.

Context-Consistent Stitching. We here investigate the effectiveness of three elements of CCS implementation: (a) context-consistent strategy, (b) stitching together regions from different images, and (c) increasing the number of labeled HO pairs. Results are shown in Table 3c. For (a), we randomly stitch union regions from different images without considering context consistency. It is shown that the mAP drops by 0.83%. This may be because HOI detection depends more on visual context; therefore, the synthesized images should be visually realistic. For (b), we simply stitch a selected union region from one image for four times. The mAP subsequently drops by 0.79%, which shows that diversity is also helpful in boosting performance. For (c), we try to keep the number of labeled HO pairs in one synthesized image equal to that of the original one. In this case, the mAP drops by 1.28%, showing that increasing the number of labeled HO pairs plays an important role in improving transformer's set prediction power.

Table 4. Performance comparisons on HICO-DET.

	Method	Detector	Backbone	Default Mode full	rare	non-rare
Two-Stage	SG2HOI [18]	COCO	ResNet-50	20.93	18.24	21.78
	DJ-RN [33]	COCO	ResNet-50	21.34	18.53	22.18
	SCG [20]	COCO	ResNet-50-FPN	21.85	18.11	22.97
	ConsNet [39]	COCO	ResNet-50-FPN	22.15	17.12	23.65
	PastaNet [35]	COCO	ResNet-50	22.65	21.17	23.09
	IDN [36]	COCO	ResNet-50	23.36	22.47	23.63
	DRG [47]	HICO-DET	ResNet-50-FPN	24.53	19.47	26.04
	IDN [36]	HICO-DET	ResNet-50	24.58	20.33	25.86
One-Stage	IP-Net [30]	COCO	Hourglass-104	19.56	12.79	21.58
	HOTR [26]	COCO	ResNet-50	23.46	16.21	25.62
	ASNet [24]	COCO	ResNet-50	24.40	22.39	25.01
	GGNet [22]	HICO-DET	Hourglass-104	23.47	16.48	25.60
	PST [19]	HICO-DET	ResNet-50	23.93	14.98	26.60
	HOI-Trans [25]	HICO-DET	ResNet-101	26.61	19.15	28.84
	ASNet [24]	HICO-DET	ResNet-50	28.87	24.25	30.25
	QPIC [23]	HICO-DET	ResNet-50	29.07	21.85	31.23
	QPIC [23]	HICO-DET	ResNet-101	29.90	23.92	31.69
	CND-S [37]	HICO-DET	ResNet-50	31.44	27.39	32.64
	Ours (HOTR)	COCO	ResNet-50	**25.97**	**26.09**	**25.93**
	Ours (QPIC)	HICO-DET	ResNet-50	**31.55**	**26.75**	**32.99**
	Ours (QPIC)	HICO-DET	ResNet-101	**31.80**	**25.95**	**33.55**
	Ours (CDN-S)	HICO-DET	ResNet-50	**33.28**	**29.19**	**34.50**

Table 5. Performance comparisons on HOI-A.

	Method	Backbone	mAP
Two-Stage	iCAN [16]	ResNet-50	44.23
	TIN [17]	ResNet-50	48.64
	GMVM [62]	ResNet-50	60.26
	C-HOI [31]	ResNet-50	66.04
One-Stage	PPDM [29]	Hourglass-104	71.23
	AS-Net [24]	ResNet-50	72.19
	QPIC [23]	ResNet-50	74.10
	Ours (QPIC)	ResNet-50	**76.87**

4.5. Comparisons with State-of-the-Art Methods

Performance Comparisons. As shown in Table 4, our approach outperforms all state-of-the-art two-stage and one-stage methods on HICO-DET. More specifically, it outperforms the QPIC baseline by 2.48%, 4.90% and 1.76% mAP in DT mode for the full, rare and non-rare HOI categories respectively with ResNet-50 as backbone. Further, we apply our approach on HOTR [26] and CDN [37]. The results further show that our method with the ResNet-50 backbone achieves a 1.84% (2.51%) mAP performance gain in DT mode for the full HOI categories over the CDN (HOTR) baseline. The complete results for both DT and KO mode are presented in the supplementary material.

Furthermore, the results on HOI-A are presented in Table 5. Our method can be seen to outperform all state-of-the-art methods by significant margins. In particular, our method

Table 6. Performance comparisons on V-COCO.

	Method	Backbone	mAP_{role}
Two-Stage	VSGNet [32]	ResNet-152	51.8
	FCL [27]	ResNet-50-FPN	52.3
	SCG [20]	ResNet-50-FPN	53.0
	FCMNet [45]	ResNet-50	53.1
	SG2HOI [18]	ResNet-50	53.3
One-Stage	HOI-Trans [25]	ResNet-101	52.9
	ASNet [24]	ResNet-50	53.9
	GGNet [22]	Hourglass-104	54.7
	HOTR [26]	ResNet-50	55.2
	DIRV [38]	EfficientDet-d3	56.1
	QPIC [23]	ResNet-50	58.8
	CDN-S [37]	ResNet-50	61.7
	Ours (QPIC)	ResNet-50	**63.5**

outperforms QPIC by a large margin of 2.77% in terms of mAP when the same backbone is used.

We present the comparison results on V-COCO in Table 6. It is found that our approach outperforms all other methods, achieving 63.5% in terms of mAP_{role} on Scenario 1.

Training Efficiency. In Figure 1, it is shown that our method significantly accelerates the training convergence of QPIC. More specifically, our approach significantly reduces the number of training epochs required on the HICO-DET dataset. For example, when the number reduces to one third of that for QPIC, our approach still achieves better performance. We also apply our method to HOTR [26] and CDN [37]. Improvements in both training efficiency and detection accuracy on both models are achieved (see supplementary material for further details).

Table 7. Comparisons in different spatial priors on QPIC. Experiments are conducted in DT mode of HICO-DET.

Spatial Priors	Full	Rare	Non-rare	Epoch
-	29.07	21.85	31.23	120
ROI	29.24	22.18	31.35	100
Gaussian map	29.58	24.11	31.21	100
Ours	**30.41**	**25.10**	**32.00**	**80**

There are several works [12, 14] that focus on accelerating the training convergence of the DETR model for object detection. One popular strategy is to impose spatial priors regarding object locations on the attention maps of each decoder layer; these priors can be a Gaussian-like distribution map [12] or a Region of Interest (ROI) [14]. Here, we apply both of these approaches to QPIC. Specifically, in the first model, we generate a 2D Gaussian-like weight map for each HOI query according to the estimated location of the human-object union region by the former decoder layer. The weight map is added to the attention maps in cross-attention in an element-wise manner. In the second model, we perform cross-attention only for pixels within the estimated human-object union regions for each HOI query. As shown in Table 7, DOQ achieves the best performance, which indicates that this method is more suitable for HOI detection than the other two methods. This may be because HOI detection is more strongly dependent on visual context, while the other two methods may suppress visual context through strong priors regarding object locations. By contrast, DOQ detects important visual context using oracle queries. By prompting the student model to attend to the same visual context, both training convergence and detection performance are improved.

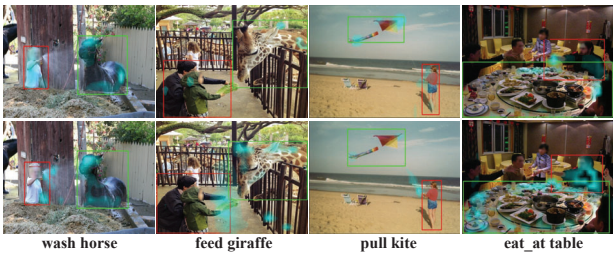

Figure 5. Visualization of HOI detection results and attention maps in decoder layers. The two rows represent results for QPIC and our method, respectively.

4.6. Qualitative Comparisons

We further visualize the attention maps produced by cross-attention in the transformer decoder layers of both QPIC [23] (the first row) and our model (the second row) in Figure 5. Here, images are randomly selected form the test set of HICO-DET database. We can observe that our model produces more reasonable attention maps, which can highlight important pixels from across the whole image for interaction category prediction. More qualitative comparisons results are provided in the supplementary material.

5. Conclusion and Limitations

In this paper, we aim to address the problems of constrained representation learning ability and constrained set prediction power for transformer-based HOI detection methods. First, we propose a knowledge distillation model that utilizes oracle HOI queries to provide additional supervision. Second, we introduce an efficient data augmentation method that synthesizes new images with more labeled HO pairs. We conduct a series of experiments on three benchmarks to demonstrate the superiority of our methods. Our work also has certain limitations. For example, we have not yet addressed the problem of the large memory costs incurred due to the self- and cross-attention operations in transformer. In the future, we will explore ways to build a less memory-intensive approach to transformer-based HOI detection.

Broader Impacts. HOI detection predicts the position of human and objects in an image and infers their interactions, which means that it is useful in various real-world applications, e.g., health care system and autonomous driving. Moreover, to the best of our knowledge, our work does not have obvious negative social impacts.

Acknowledgement. This work was supported by National Natural Science Foundation of China under Grant 62076101, the Program for Guangdong Introducing Innovative and Entrepreneurial Teams under Grant 2017ZT07X183, and CCF-Baidu Open Fund.

References

[1] H. Zhao, R. Wildes. Spatiotemporal feature residual propagation for action prediction. In *ICCV*, 2019. 1

[2] Y. Kong, Z. Tao, Y. Fu. Deep sequential context networks for action prediction. In *CVPR*, 2017 1

[3] X. Lin, C. Ding, J. Zeng, D. Tao. Gps-net: Graph property sensing network for scene graph generation. In *CVPR*, 2020. 1

[4] M. Suhail, A. Mittal, B. Siddiquie, C. Broaddus, J. Eledath, G. Medioni, L. Sigal. Energy-Based Learning for Scene Graph Generation. In *CVPR*, 2021. 1

[5] K. Tang, H. Zhang, B. Wu, W. Luo, W. Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, 2019. 1

[6] L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, Y. Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *CVPR*, 2020. 1

[7] J. Yim, D. Joo, J. Bae, J. Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, 2017. 5

[8] F. Tung, G. Mori. Similarity-preserving knowledge distillation. In *ICCV*, 2019. 5

[9] P. Chen, S. Liu, H. Zhao, J. Jia. Distilling Knowledge via Knowledge Review. In *CVPR*, 2021. 5

[10] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *ICLR*, 2020. 3

[11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2, 3, 4, 6

[12] P. Gao, M. Zheng, X. Wang, J. Dai, H. Li. Fast Convergence of DETR With Spatially Modulated Co-Attention. In *ICCV*, 2021. 3, 8

[13] D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, L. Sun, J. Wang. Conditional DETR for Fast Training Convergence. In *ICCV*, 2021. 3, 4

[14] X. Dai, Y. Chen, J. Yang, P. Zhang, L. Yuan, L. Zhang. Dynamic DETR: End-to-End Object Detection With Dynamic Attention. In *ICCV*, 2021. 3, 8

[15] X. Jiang, Z. Chen, Z. Wang, E. Zhou, C. Yuan. Guiding Query Position and Performing Similar Attention for Transformer-Based Detection Heads. arXiv:2108.09691, 2021. 3

[16] C. Gao, Y. Zou, J. Huang. iCAN: Instance-Centric Attention Network for Human-Object Interaction Detection. In *BMVC*, 2018. 7

[17] Y. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H. Fang, Y. Wang, C. Lu. Transferable interactiveness knowledge for human-object interaction detection. In *CVPR*, 2019. 7

[18] T. He, L. Gao, J. Song, Y. Li. Exploiting Scene Graphs for Human-Object Interaction Detection. In *ICCV*, 2021. 7, 8

[19] Q. Dong, Z. Tu, H. Liao, Y. Zhang, V. Mahadevan, S. Soatto. Visual Relationship Detection Using Part-and-Sum Transformers With Composite Queries. In *ICCV*, 2021. 2, 7

[20] F. Zhang, D. Campbell, S. Gould. Spatially Conditioned Graphs for Detecting Human-Object Interactions. In *ICCV*, 2021. 7, 8

[21] S. Wang, K. Yap, H. Ding, J. Wu, J. Yuan, Y. Tan. Discovering Human Interactions With Large-Vocabulary Objects via Query and Multi-Scale Detection. In *ICCV*, 2021. 2

[22] X. Zhong, X. Qu, C. Ding, D. Tao. Glance and Gaze: Inferring Action-aware Points for One-Stage Human-Object Interaction Detection. In *CVPR*, 2021. 2, 7, 8

[23] M. Tamura, H. Ohashi, T. Yoshinaga. QPIC: Query-Based Pairwise Human-Object Interaction Detection with Image-Wide Contextual Information. In *CVPR*, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[24] M. Chen, Y. Liao, S. Liu, Z. Chen, F. Wang, C. Qian. Reformulating hoi detection as adaptive set prediction. In *CVPR*, 2021. 1, 2, 3, 7, 8

[25] C. Zou, B. Wang, Y. Hu, J. Liu, Q. Wu, Y. Zhao, B. Li, C. Zhang, C. Zhang, Y. Wei, J. Sun. End-to-End Human Object Interaction Detection with HOI Transformer. In *CVPR*, 2021. 1, 2, 3, 4, 7, 8

[26] B. Kim, J. Lee, J. Kang, E. Kim, H. Kim. HOTR: End-to-End Human-Object Interaction Detection with Transformers. In *CVPR*, 2021. 1, 2, 3, 4, 6, 7, 8

[27] Z. Hou, B. Yu, Y. Qiao, X. Peng, D. Tao. Detecting human-object interaction via fabricated compositional learning. In *CVPR*, 2021. 8

[28] Z. Hou, B. Yu, Y. Qiao, X. Peng, D. Tao. Affordance Transfer Learning for Human-Object Interaction Detection. In *CVPR*, 2021. 2

[29] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, J. Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *CVPR*, 2020. 2, 5, 6, 7

[30] T. Wang, T. Yang, M. Danelljan, F. Khan, X. Zhang, J. Sun. Learning human-object interaction detection using interaction points. In *CVPR*, 2020. 2, 7

[31] T. Zhou, W. Wang, S. Qi, H. Ling, J. Shen. Cascaded human-object interaction recognition. In *CVPR*, 2020. 2, 7

[32] O. Ulutan, A. Iftekhar, B. Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *CVPR*, 2020. 2, 8

[33] Y. Li, X. Liu, H. Lu, S. Wang, J. Liu, J. Li, C. Lu. Detailed 2d-3d joint representation for human-object interaction. In *CVPR*, 2020. 2, 7

[34] S. Wang, K. Yap, J. Yuan, Y. Tan. Discovering human interactions with novel objects via zero-shot learning. In *CVPR*, 2020. 2

[35] Y. Li, L. Xu, X. Liu, X. Huang, Y. Xu, S. Wang, H. Fang, Z. Ma, M. Chen, C. Lu. Pastanet: Toward human activity knowledge engine. In *CVPR*, 2020. 2, 7

[36] Y. Li, X. Liu, X. Wu, Y. Li, C. Lu. HOI Analysis: Integrating and Decomposing Human-Object Interaction. In *NeurIPS*, 2020. 7

[37] A. Zhang, Y. Liao, S. Liu, M. Lu, Y. Wang, C. Gao, X. Li. Mining the Benefits of Two-stage and One-stage HOI Detection. In *NeurIPS*, 2021. 3, 6, 7, 8

[38] H. Fang, Y. Xie, D. Shao, C. Lu. DIRV: Dense Interaction Region Voting for End-to-End Human-Object Interaction Detection. In *AAAI*, 2021. 8

[39] Y. Liu, J. Yuan, C. Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *ACM MM*, 2020. 2, 7

[40] B. Kim, T. Choi, J. Kang, H. Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *ECCV*, 2020. 2

[41] Z. Hou, X. Peng, Y. Qiao, D. Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, 2020. 2

[42] X. Zhong, C. Ding, X. Qu, D. Tao. Polysemy deciphering network for human-object interaction detection. In *ECCV*, 2020. 2, 4

[43] X. Zhong, C. Ding, X. Qu, D. Tao. Polysemy Deciphering Network for Robust Human–Object Interaction Detection. In *IJCV*, 2021. 2

[44] H. Wang, W. Zheng, L. Yingbiao. Contextual heterogeneous graph network for human-object interaction detection. In *ECCV*, 2020. 2

[45] Y. Liu, Q. Chen, A. Zisserman. Amplifying key cues for human-object-interaction detection. In *ECCV*, 2020. 2, 8

[46] D. Kim, X. Sun, J. Choi, S. Lin, I. Kweon. Detecting human-object interactions with action co-occurrence priors. In *ECCV*, 2020. 2

[47] C. Gao, J. Xu, Y. Zou, J. Huang. Drg: Dual relation graph for human-object interaction detection. In *ECCV*, 2020. 7

[48] C. Yu-Wei, L. Yunfan, L. Xieyang, Z. Huayi, D. Jia. Learning to Detect Human-Object Interactions. In *WACV*, 2018. 1, 2, 5, 6

[49] S. Gupta, J. Malik. Visual Semantic Role Labeling. arXiv:1505.04474, 2015. 2, 6

[50] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, Microsoft COCO: Common objects in context. In *ECCV*, 2014. 6

[51] K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[52] I. Loshchilov, F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[53] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, 2013. 7

[54] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 6, 7

[55] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba. Places: A 10 million Image Database for Scene Recognition. In *IEEE TPAMI*, 2017. 5

[56] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T. Lin, E. Cubuk, Q. Le, B. Zoph. Simple Copy-Paste Is a Strong Data Augmentation Method for Instance Segmentation. In *CVPR*, 2021. 5

[57] D. Dwibedi, I. Misra, M. Hebert. Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection. In *ICCV*, 2017. 5

[58] N. Dvornik, J. Mairal, C. Schmid. Modeling Visual Context is Key to Augmenting Object Detection Datasets. In *ECCV*, 2018. 5

[59] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 5

[60] T. Lin, P. Goyal, R. Girshick, K. He, P. Dollar. Focal loss for dense object detection. In *ICCV*, 2017. 5

[61] H. Kuhn. The Hungarian method for the assignment problem. In *Naval research logistics quarterly*, 1955. 4

[62] Pic leaderboard. http://www.picdataset.com/challenge/leaderboard/hoi2019, 2019. 7