What to look at and where: Semantic and Spatial Refined Transformer for detecting human-object interactions

A S M Iftekhar*†, Hao Chen*‡, Kaustav Kundu‡, Xinyu Li‡, Joseph Tighe‡, Davide Modolo‡
‡AWS AI Labs; †University of California, Santa Barbara
{hxen,kaustavk,-,tighej,dmodolo}@amazon.com; iftekhar@ucsb.edu

Abstract

We propose a novel one-stage Transformer-based semantic and spatial refined transformer (SSRT) to solve the Human-Object Interaction detection task, which requires to localize humans and objects, and predicts their interactions. Differently from previous Transformer-based HOI approaches, which mostly focus at improving the design of the decoder outputs for the final detection, SSRT introduces two new modules to help select the most relevant object-action pairs within an image and refine the queries' representation using rich semantic and spatial features. These enhancements lead to state-of-the-art results on the two most popular HOI benchmarks: V-COCO and HICO-DET.

1. Introduction

Human-object interaction (HOI) detection is an important building block for complex visual reasoning, such as scene understanding [11, 51] and action recognition [50, 55], and its goal is to detect all HOI triplets $\langle human, object, action \rangle$ in each image. Fig. 1 shows an example of a HOI detection, where the *person* (i.e., the human) is denoted with a red bounding box, the *sports ball* (i.e., the object) with a yellow bounding box, and the action *kick* is what that human is performing with that object.

The HOI literature can be divided into *two-stage* and *one stage* approaches. *Two-stage* approaches [12–14, 18, 19, 27, 30, 31, 33, 34, 39, 43, 46–48, 53, 54, 56] first use off-the-shelf detectors to localize all instances of people and objects independently. For each person and object bounding box pair, an interaction class is then predicted in the second stage. This sequential process has two main drawbacks [7, 25, 26]: (1) off-the-shelf object detectors are agnostic to the concept of interactions; and (2) enumerating over all pairs of person and object bounding boxes to predict an interaction class is time-consuming and expensive. In contrast, *one-stage* approaches detect all the components of an HOI triplet directly in an end-to-end fashion. Some earlier one-stage approaches used intermediate representations based on interaction points [32, 49] and union boxes [25] to predict these. However, such methods fail when the interacting human and object are far away from each other and when multiple interactions overlap (e.g., crowd scenes) [7, 42].

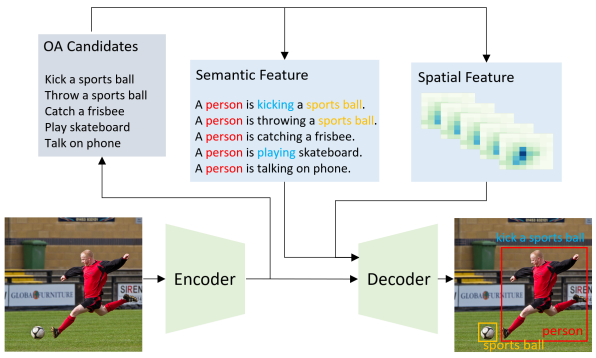


Figure 1. *Conceptual workflow of SSRT. Instead of just feeding the encoded image to the decoder, we pre-select object-action (OA) prediction candidates and encode them to semantic and spatial features. These features then refine the learnt queries in decoding to enable them to attend to more relevant HOI predictions.*

More recently, a new trend of one-stage approaches [7, 26, 42, 58] based on *Transformer architectures* [5, 10, 35] have been proposed to overcome these problems and improve the HOI detection performance. This paper belongs to this category of works (Fig. 1). At a high-level, these approaches first use a CNN backbone to extract image features and then feed them into an encoder-decoder architecture. Some approaches use two decoders to detect instances and interactions in parallel [7, 26]. while others follow a simpler design that directly predicts all the elements of an HOI triplet with a single decoder [42, 58]. While successful, this design suffers from two limitations: (i) not all object-action pairs are meaningful (e.g., a person cannot be 'cutting a pizza' when the pizza is far away from the person's location; and it is unusual for a person to be 'cutting a football'),

*Equal contribution.
†Work done during an internship at Amazon.

simply relying on the one-shot network to reduce them may not be effective; and (ii) each simple query is decoded for all the rich elements of an HOI triplet (i.e., person location, object class and location, and interaction class), which is challenging, especially considering how HOI detection requires reasoning about complex relational structures in images. We address both of these limitations in our work.

For this, we propose **S**emantic and **S**patial **R**efined **T**ransformer (SSRT), that solves the aforementioned limitations by predicting what subset of object-action pairs is relevant for an image, and using explicit semantic and spatial information to support and guide the queries, so that they can be decoded more reliably and are more aligned with the final detection. In details, SSRT improves the Transformer design of previous HOI detector by introducing two novel modules: a Support Feature Generator (SFG) and a Query Refiner (QR) (Fig. 2). The former generates semantic and spatial features from a set of pre-selected *object-action (OA)* pairs, while the latter integrates these features for decoding.

Our approach achieves state-of-the-art results on both V-COCO [16] and HICO-DET [6] datasets, showing the importance and effectiveness of our semantic and spatial guidance for HOI detection. Finally, in an extensive ablation study, we also evaluate our model design and our parameter choices, to further highlight the SSRT's contributions.

2. Related Work

Two stage HOI detection networks detect objects and then detect HOIs among those detected objects. These networks rely on an off-the-shelf object detector [41] to localize objects. For detecting interactions among the detected objects these networks develop different novel techniques. Few works [13,34] consider humans and objects as nodes in graph networks. Another line of works [43, 46] utilize spatial and pose features to attend salient spatial regions of the images. Additionally, other works are using object affordance based architectures [20, 21] to deal with the long-tail distribution problem of the HOI detection datasets. Moreover, there are works [19, 30] that leverage the compositional nature of objects and interactions to detect HOIs. Another paradigm of two stage works utilize additional features like 3D representation of humans [29], semantic contexts [22, 34], segmentation masks [33]. However, performance of these networks are highly dependent on the quality of the object detection. Moreover, these networks suffer heavily to process the overwhelming number of non-interacting detected objects [57].

To deal with the issues faced by two stage networks, recent works [7, 9, 26, 42, 57, 58] are trying to detect HOIs in a one stage framework. These networks take images as input and directly detect and localize HOIs over those images. Initial one stage HOI detection networks [32, 49] focus on detecting pre-defined interaction points to detect interactions. However, these heuristic based approaches often fail to find spatial contextual information. For getting richer contextual features many recent one stage HOI detection networks [7, 26, 42, 58] adapt encoder-decoder based Transformer [45] like architecture inspired from the one stage object detection network DETR [5].

However, these networks do not consider the additional complexity of doing two related but different subtasks of object localization and interaction detection. The base network of these mentioned works is essentially an object detector network which is expanded for interaction detection. Therefore, it is beneficial to provide additional guidance to these networks. Moreover, these one stage networks do not leverage spatial and semantic cues that are proven to be beneficial to detect HOIs in few two-stage works [22, 44, 52]. In this respect, we propose a semantic and spatially refined transformer based architecture to detect HOIs in one single stage. Our superior numerical results over these state of the art methods prove our method's effectiveness.

3. Technical Approach

Most of today's Transformer-based HOI detection networks [7, 26, 42, 58] follow the DETR [5] architecture and focus on improving the *design of the decoder outputs* for the HOI task. Instead, our SSRT approach improves the overall *design of the Transformer*. Specifically, it adds two new modules between the encoder and the decoder: a Support Feature Generator (**SFG**) (Sec. 3.2) and a Query Refiner (**QR**) (Sec. 3.3) (Fig. 2). At a high level, SSRT works as follows: given an input image, it first extracts its features with a CNN backbone and then transform those using a transformer encoder. Instead of feeding the encoded features directly into the decoder, the features are sent to SFG to first generate a set of *object-action (OA)* prediction candidates (without localization). Then the SFG generates both spatial and semantic features using these candidates and aggregates them as support features. These support features are then sent to the QR to refine the learnable queries. Finally, the HOI Decoder takes the inputs as both the encoded features and the refined queries, and outputs a set of embeddings, each of which is used to predict an HOI output.

3.1. Our Architecture

Given an input image $\mathbf{x} \in \mathbb{R}^{H_0 \times W_0 \times C_0}$, where H_0, W_0, C_0 denote the image height, width and color channels, SSRT first extracts a feature map $\mathbb{R}^{H \times W \times C}$ using a CNN backbone ($\mathbf{F}$) (e.g. ResNet-50 [17]). $\mathbf{F}(\mathbf{x})$ is then sent to a 1x1 convolution to reduce the channel dimension C to a smaller value d, and obtain $\mathbf{F}_c(\mathbf{x}) \in \mathbb{R}^{H \times W \times d}$. Following previous works [7, 26, 42, 58], we add a fixed positional encoding $\mathbf{p} \in \mathbb{R}^{H \times W \times d}$ to the input feature of the encoder to supplement the positional information. The encoder follows the standard architecture of the transformer as a stack

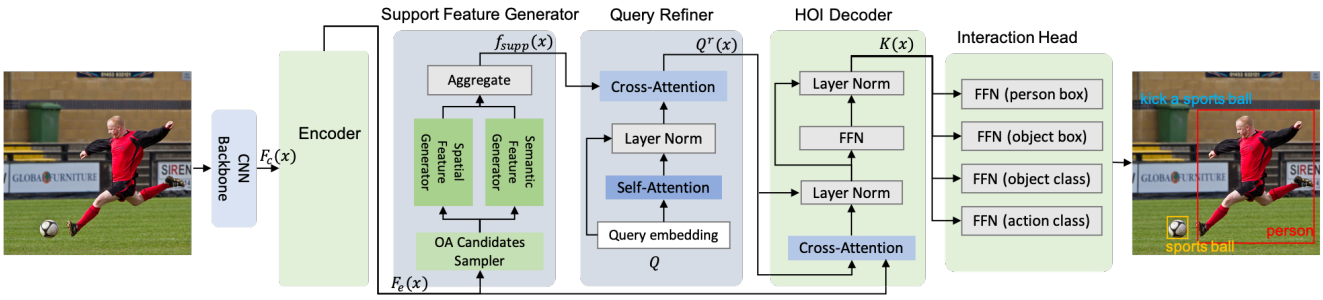


Figure 2. *Overview of our SSRT network. Given an image, we extract features using a backbone and feed them to Transformer Encoder. The encoder's output is then sent to the Support Feature Generator (SFG), which first predicts and selects top-K object-action (OA) candidates, and then generates spatial and semantic features. Next, the aggregated features are sent to the Query Refiner (QR) to refine queries. Finally, The refined queries are decoded: each query is used to predict a human bounding box, an object bounding box, an interaction vector and an object class vector via interaction heads with small FFNs.*

of multi-head self-attention modules and a feed forward network (FFN). The encoded feature map $\mathbf{F}_e(\mathbf{x}) \in \mathbb{R}^{H\times W\times d}$ is obtained as follows:

$$\mathbf{F}_e(\mathbf{x}) = \text{Encoder}(\mathbf{F}_c(\mathbf{x}), \mathbf{p}) \tag{1}$$

Instead of feeding $\mathbf{F}_e(\mathbf{x})$ only to the HOI decoder, we also send it to our SFG module. Here, we first predict K pairs of (object, action) categories (i.e., OA pairs) present in the image and select a subset from them. We then use spatial and semantic cues for the OA prediction candidates to generate support features, $\mathbf{f}_{supp}(\mathbf{x})$. The details of the SFG module are discussed in Sec. 3.2. The support features, $\mathbf{f}_{supp}(\mathbf{x})$ along with an initial set of queries, $\mathbf{Q} = \{\mathbf{q_i}|\mathbf{q_i} \in \mathbb{R}^d\}_{i=1}^{N_q}$ are fed into the QR module. The query refiner module is a decoder like architecture which outputs a refined set of queries, $\mathbf{Q}^r(\mathbf{x}) = \{\mathbf{q}_\mathbf{i}^\mathbf{r}(\mathbf{x})|\mathbf{q}_\mathbf{i}^\mathbf{r}(\mathbf{x}) \in \mathbb{R}^d\}_{i=1}^{N_q}$, i.e.,

$$\mathbf{Q}^r(\mathbf{x}) = \text{QR}(\mathbf{f}_{supp}(\mathbf{x}), \mathbf{Q}) \tag{2}$$

The details of the QR block are discussed in Sec. 3.3. Both the encoder output $\mathbf{F}_e(\mathbf{x})$ and the refined queries $\mathbf{Q}^r(\mathbf{x})$ are sent to the HOI decoder for the final decoding. Note that in contrast to the standard transformer architectures, the queries which are fed into the decoder are a function of the input image $\mathbf{x}$. The goal of modeling it in this manner is to explicitly provide more guidance to the decoder so that it can generate more accurate HOI outputs. The HOI decoder follows the standard architecture of the transformer as a stack of multi-headed cross-attention units but no self-attention layers. The refined queries $\mathbf{Q}^r(\mathbf{x})$ are transformed into a set of output embeddings, $\mathbf{K}(\mathbf{x}) = \{\mathbf{k}_i(\mathbf{x})|\mathbf{k}_i(\mathbf{x}) \in \mathbb{R}^d\}_{i=1}^{N_q}$, i.e.,:

$$\mathbf{K}(\mathbf{x}) = \text{Decoder}(\mathbf{F}_e(\mathbf{x}), \mathbf{p}, \mathbf{Q}^r(\mathbf{x})) \tag{3}$$

where $\mathbf{p}$ is the positional embedding. Each query is designed to capture at most one HOI prediction. We feed these queries to four small Feed Forward Network (FFNs) to predict human bounding boxes $\mathbf{b}_h(\mathbf{x}) \in [0,1]^4$, object bounding boxes $\mathbf{b}_o(\mathbf{x}) \in [0,1]^4$, interaction prediction vectors $\mathbf{P}_{HOI}(\mathbf{x}) \in [0,1]^{N_{act}}$, and object class prediction vectors $\mathbf{P}_{obj}(\mathbf{x}) \in [0,1]^{N_{obj}}$, where N_{act} and N_{obj} are the number of interaction classes and number of object classes. $\mathbf{b}_o$, $\mathbf{b}_h$ and $\mathbf{P}_{HOI}$ are predicted with sigmoid functions, $\mathbf{P}_{obj}$ is predicted with softmax function. Like [42], we weigh our final interaction prediction vectors with the most confident object class predictions as:

$$\mathbf{P}_{HOI}(\mathbf{x}) = \mathbf{P}_{HOI}(\mathbf{x}) * max(\mathbf{P}_{obj}(\mathbf{x})) \tag{4}$$

We discuss details on how to train this network in Sec. 3.4.

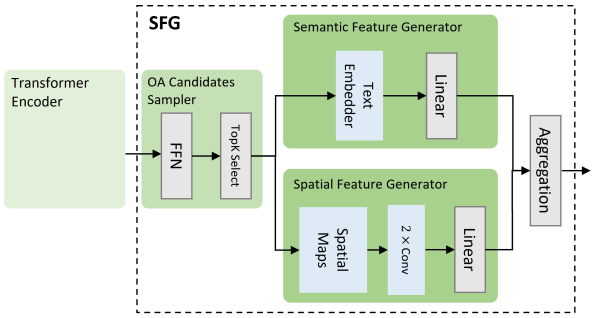


Figure 3. *Our Support Feature Generator (SFG) design.*

3.2. Support Feature Generator

The goal of SFG is to provide support to the transformer with additional semantic and spatial cues, as they play significant roles towards detecting all the rich HOI outputs. Specifically, semantic cues are important to help capture the human-object relations [52] while spatial cues are important to help accurately localize the humans and objects [43].

While explicitly using these cues have been proved successful in the two-stage solutions [43,52], it has not been exploited in one-stage approaches. In this block, we propose

how to generate support features $\mathbf{f}_{supp}(\mathbf{x})$ for an image, which can then be used as inputs to the query refiner block and subsequently to the decoder. We first select K high confident *object-action (OA)* candidates predicted from the encoder feature, $\mathbf{F}_e(\mathbf{x})$, and enrich them with semantic and spatial embeddings to generate the support features (Fig. 3).

OA Candidates Sampler. As shown in Fig. 3, we build a 3-layer FFN g_{cls} on top of the average-pooled encoded features $\mathbf{F}_e(\mathbf{x})$ to predict the *object-action (OA)* candidates, i.e., $\mathbf{s}(\mathbf{x}) = \sigma(g_{cls}(\text{avg-pool}(\mathbf{F}_e(\mathbf{x}))))$, where $\mathbf{s}(\mathbf{x}) \in [0,1]^{N_s}$, N_s is the number of possible set of object-action (OA) pairs and σ is the sigmoid function. Note how in this module, $\mathbf{s}(\mathbf{x})$ corresponds to the OA labels without localization. Among all predictions, we then select the top-K (object, action) candidates with the highest confidence. Let this set of selected candidates be represented by $\mathbf{S}_{\text{cand}} = \{(y_{o,i}, y_{a,i})\}_{i=1}^{K}$.

Semantic Feature Generator. Recently, language-image models [23,40] have shown strong capabilities in generating high quality representations that can capture rich semantic and context information. We believe such embedding should be able to capture the relational structure and enrich the context of the transformer network. Therefore, we use a CLIP [40] text encoder to compute the semantic representation of each pre-selected OA candidate. Since CLIP works best with sentences (as opposed to single words), we convert each predicted OA into a full sentence before feeding it to the CLIP text encoder. For example, we transform the pair $(phone, talk)$ into the sentence "A person is talking on the phone". The transformation is done automatically following certain pre-defined rules using scripts with tiny manual efforts. Finally we project these semantic features to the image feature space by using a linear projection layer (Fig. 3, top). For each OA candidate, $(y_{o,i}, y_{a,i}) \in \mathbf{S}_{\text{cand}}$, we compute the semantic feature as follows:

$$\mathbf{f}_{sem}(y_{o,i}, y_{a,i}) = \text{Emb}_{sem}(y_{o,i}, y_{a,i}) \tag{5}$$

where, Emb_{sem} is the semantic embedding function.

Spatial Feature Generator. In two-stage HOI approaches, the spatial features are generated based on predicted human and object bounding boxes from off-the-shelf detectors [14, 43]. As these are not available for one-state approaches like ours, we propose to estimate bounding box locations using training data statistics.

We define the *relative spatial configuration (RSC)* as the object bounding box location with respect to the human bounding box location, and estimate the RSC from the training data. Specifically, given a human h and an object o, we denote the human bounding box as (x_h, y_h, w_h, h_h) and object bounding box as (x_o, y_o, w_o, h_o), where (x, y) is the top left point and (w, h) is width and height of the bounding box. Inspired by previous work [15], we define the RSC as $(\Delta x_{oh}, \Delta y_{oh}, \Delta w_{oh}, \Delta h_{oh})$, where: $\Delta x_{oh} = \frac{x_o - x_h}{w_h}$, $\Delta y_{oh} = \frac{y_o - y_h}{h_h}$, $\Delta w_{oh} = \log \frac{w_o}{w_h}$, and $\Delta h_{oh} = \log \frac{h_o}{h_h}$. We then consider for each interaction, $\Delta x_{oh}, \Delta y_{oh}$ follow a bi-variate Gaussian distribution and $\Delta w_{oh}, \Delta h_{oh}$ follow another bi-variate Gaussian distribution. We estimate essential parameters (mean, co-variance) for these variables using all the training samples for each OA label. We estimate the person bounding boxes in the similar way.

Using these distributions, we then generate random samples as the human and object bounding boxes to create the spatial features. As shown in Fig. 3 we follow previous works [43] to generate the spatial map for each OA label. The spatial map is a $2 \times B \times B$ size binary map where in the first channel the location of the human bounding box is 1 and in the second channel the location of the object bounding box is 1. The rest of the locations in spatial map are zero. Finally we pass through this spatial map to 2 convolution layers followed by a linear projection layer to generate the spatial feature. For each OA candidate, $(y_{o,i}, y_{a,i}) \in \mathbf{S}_{\text{cand}}$, we compute the spatial features:

$$\mathbf{f}_{spa}(y_{o,i}, y_{a,i}) = \text{Emb}_{spa}(y_{o,i}, y_{a,i}) \tag{6}$$

Emb_{spa} is the embedding function for the spatial features.

Feature Aggregation. For each pre-selected OA candidate, $(y_{o,i}, y_{a,i}) \in \mathbf{S}_{\text{cand}}$, we have the semantic feature, $\mathbf{f}_{sem}(y_{o,i}, y_{a,i})$ and the spatial feature, $\mathbf{f}_{spa}(y_{o,i}, y_{a,i})$. These features are aggregated as follows:

$$\mathbf{f}_{agr}(y_{o,i}, y_{a,i}) = g_{agr}(\mathbf{f}_{sem}(y_{o,i}, y_{a,i}), \mathbf{f}_{spa}(y_{o,i}, y_{a,i})) \tag{7}$$

where g_{agr} is the aggregation function. We concatenate the features, $\mathbf{f}_{agr}(y_{o,i}, y_{a,i})$ extracted for all candidates $\in \mathbf{S}_{\text{cand}}$ and form the support feature, $\mathbf{f}_{supp}(\mathbf{x}) \in \mathbb{R}^{K \times d}$.

3.3. Query Refiner

The query refiner is designed to use the pre-selected OA candidates and support features generated from SFG module to refine the learnt queries that are randomly initialized. Ideally these pre-generated contextual signals should be able to guide the queries to be learnt to attend to more relevant candidates and reduce noisy predictions. To achieve this we cross-attend the learnt queries with support features.

Specifically, as shown in Fig. 2, the query refiner is built on standard transformer decoder structure. The randomly initialized queries $\mathbf{Q} = \{\mathbf{q_i} | \mathbf{q_i} \in \mathbb{R}^d\}_{i=1}^{N_q}$ first attend to themselves via self-attention. Then these queries attend to support features $\mathbf{f}_{supp}(\mathbf{x})$ generated from the SFG (Sec. 3.2) through cross-attention. Here, the support features serve as keys and values to the attention architecture. As a result, queries have additional direction to look for correct object-action in the encoded image features. In the final HOI decoder, queries attend to the encoded image features. The output of the decoder are the context-aware queries which contain rich cues to detect HOIs.

3.4. Training details

To train this network, we apply the same loss function as [42] at the outputs of the interaction head, $\mathbf{b}_o$, $\mathbf{b}_h$, $\mathbf{P}_{HOI}$ and $\mathbf{P}_{obj}$. The loss calculation is composed of two stages: the bipartite matching stage between predictions and ground truths, and the loss calculation stage for the matched pairs. For the bipartite matching, we follow the training procedure of DETR [5] and use the Hungarian algorithm [28]. Then the loss is calculated on the basis of the matched pairs as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{box}} + \lambda_2 \mathcal{L}_{\text{iou}} + \lambda_3 \mathcal{L}_{\text{obj}} + \lambda_4 \mathcal{L}_{\text{HOI}}, \tag{8}$$

where $\mathcal{L}_{\text{box}}$ and $\mathcal{L}_{\text{iou}}$ are l_1 and $GIoU$ loss applied to both human and object bounding boxes, $\mathcal{L}_{\text{obj}}$ is a cross entropy loss for object prediction, and $\mathcal{L}_{\text{HOI}}$ is a binary cross entropy loss for interaction prediction. λ_1, λ_2, λ_3 and λ_4 are hyperparameters selected following [42].

Additionally we use a binary cross entropy loss for the $\mathbf{s}$ output, which corresponds to the image-level (object, action) pair prediction. All these losses are trained in a multi-task setting.

4. Experimental settings

Dataset & Metrics. We evaluate SSRT on the two most popular benchmark datasets: V-COCO [16] and HICO-DET [6]. **V-COCO** has 29 interaction classes. Following [31], we evaluate the performance on 24 interaction classes since 4 interaction classes have no object pair and 1 class has very few samples. This dataset has 2,533 training, 2,867 validation and 4,946 testing images. **HICO-DET** [6] has 600 human-object interaction classes. It consists of 38,117 training and 9,658 test images.

We report mean average precision (mAP) on the test set for both V-COCO and HICO-DET datasets. A prediction is considered to be correct if the predicted human and object bounding boxes overlap (with IoU greater than 0.5) with the respective GT boxes and the predicted interaction class is correct. We follow the protocol established in [43] to evaluate results on the V-COCO dataset. For human actions that do not interact with any object, two evaluation scenarios are considered. Scenario 1 considers a strict evaluation criteria that requires the prediction of a null bounding box with coordinates $[0, 0, 0, 0]$, Scenario 2 relaxes this condition for such cases by ignoring the predicted bounding box for evaluation. We use the protocol from [6] to evaluate on the HICO-DET. The mAP metric is computed in default settings for three categories: Full (all 600 HOI classes), Rare (138 classes that have less than 10 training samples), Non-rare (462 classes that have more than 10 training samples).

Implementation Details. The architecture design is similar to that of QPIC [42]. We use ResNet-50 and ResNet-101 backbones [17]. The parameters of the network are initialized with DETR [5] trained on the COCO dataset [42]. Each of the encoder and decoder have 6 layers and 8 heads. The dimension inside the transformer architecture is 256. The total number of queries is 100. The initial learning rate of the backbone network is 10^{-5}, with others 10^{-3}. The weight decay is 10^{-4}. The learning rate is dropped at every 65 epochs and we train 150 epochs in total. We use the AdamW [36] optimizer and the batch size is 16.

We experiment with the following semantic feature generator: (a) one-hot, (b) GLOVE [37], (c) CLIP [40]. In the spatial feature generator, we use a $2 \times 64 \times 64$ dimensional binary spatial map [14, 43]. For the human bounding box location, we select (16, 16) as the fixed top-left point, as in the evaluating HOI datasets interacting human bounding boxes are mostly confined at the top-left corner of the images [43]. Both spatial and semantic features are projected to a 256-dimensional space.

5. Results

In this section, we first compare the performance of our SSRT network with the SOTA methods in Sec. 5.1, followed by an ablation study to validate the design choices in Sec. 5.2. Finally, we show qualitative analysis in Sec. 5.3.

5.1. Comparison with SOTA

In Tables 1 and 2, we compare the performance of our SSRT model to the SOTA methods on the V-COCO [16] and HICO-DET [6] datasets respectively. We group the approaches into one-stage and two-stage. Following the literature, we report numbers of SSRT with both ResNet-50 (R-50) and ResNet-101 (R-101) backbones. Results show that our SSRT has achieved SOTA performance on both datasets with the ResNet-50 backbone, while ResNet-101 can improve the performance further. We outperform all the DETR based solutions (HOI-Trans, ASNet, HOTR and QPIC) on both datasets, and overall we achieve about 10% improvement on V-COCO and 5% improvement on HICO-DET comparing to the SOTA.

5.2. Ablation Studies

In this section, we do ablation for the different design choices of SSRT. We evaluate on V-COCO dataset with the ResNet-50 backbone. For each ablation, we change one parameter, and keep the other parameters at the best setting.

Support Feature Generator Module. In Table 3a, we explore the benefits of using semantic and spatial features to generate features for the query refiner block. Comparing to the QPIC baseline (Row 1), using semantic features (Row 2) significantly improves the performance by +3.9 points. This demonstrates the effectiveness of using the semantic information to guide the HOI detection. On top of this, we explore two different ways to aggregate the semantic and

Type	Method	Scenario 1	Scenario 2
Two Stage	VCL [19]	48.3	-
	DRG [13]	51.0	-
	Wang et al. [47]	52.3	-
	FCL [21]	52.4	-
	PD-Net [56]	52.6	-
	ACP [27]	53.0	-
	FCMNet [33]	53.1	-
	SG2HOI [18]	53.3	-
	IDN [30]	53.3	60.3
	GTNet [22]	56.2	60.1
	SABRA [24]	56.6	-
One Stage	UnionDet [25]	47.5	56.2
	Wang et al. [49]	51.0	-
	HOI-Trans [58]	52.9	-
	ASNet [7]	53.9	-
	GGNet [57]	54.7	-
	HOTR [26]	55.2	64.4
	DIRV [12]	56.1	-
	QPIC(R-50) [42]	58.8	61.0
	QPIC(R-101) [42]	58.3	60.7
	Ours (R-50)	63.7	65.9
	Ours (R-101)	**65.0**	**67.1**

Table 1. *Performance comparisons on the V-COCO [16] test set. Best result is marked with **bold** and the second best result is marked with underline.*

Type	Method	Full	Rare	Non-rare
Two Stage	Wang et al. [47]	17.57	16.85	17.78
	FCMNet [33]	20.41	17.34	21.56
	ACP [27]	20.59	15.92	21.98
	PD-Net [56]	20.81	15.90	22.28
	SG2HOI [18]	20.93	18.24	21.78
	VCL [19]	23.63	17.21	25.55
	DRG [13]	24.53	19.47	26.04
	SABRA [24]	26.09	16.29	29.02
	IDN [30]	26.29	22.61	27.39
	GTNet [22]	26.78	21.02	28.50
	ATL [20]	28.53	21.64	30.59
	FCL [21]	29.12	23.67	30.75
One Stage	UnionDet [25]	17.58	11.72	19.33
	Wang et al. [49]	19.56	12.79	21.58
	PPDM [32]	21.73	13.78	24.10
	DIRV [12]	21.78	16.38	23.39
	HOI-Trans [58]	23.46	16.91	25.41
	PST [9]	23.93	14.98	26.60
	HOTR [26]	25.10	17.34	27.42
	ASNet [7]	28.87	24.25	30.25
	GGNet [57]	29.17	22.13	30.84
	QPIC(R-50) [42]	29.07	21.85	31.23
	QPIC(R-101) [42]	29.90	23.92	31.69
	Ours (R-50)	30.36	**25.42**	31.83
	Ours (R-101)	**31.34**	24.31	**33.32**

Table 2. *Performance comparisons on the HICO-DET [6] test set. Best result is marked with **bold** and the second best result is marked with underline.*

spatial features: (1) concatenation (Row 3); and (2) elementwise multiplication (Row 4). Results show that elementwise multiplication gives the best performance, which we believe is because that multiplication operates as a gating mechanism that effectively fuses semantic and spatial information, as also observed in other work [38, 43].

Semantic Inputs. Table 3b explores different kinds of semantic input that can be encoded as semantic features. For this experiment, all varieties of semantic input are encoded by the CLIP [40] text embedding model. We explore the following types of semantic input: (a) $action\ only$: using only predicted action category from the encoder. For example, if the OA prediction is $\langle laptop, work \rangle$, we only use the predicted action (i.e., "work" here) as the semantic input, (b) *object-action (OA)*: Using the previous example, the semantic input here is $\langle laptop, work \rangle$ tuple, (c) *semantic retrieval*: In this approach we model the semantic input in a non-parametric fashion. Using a joint visual-semantic embedding network [40], we retrieve nearest OA semantic tuples based on the visual features of the input. The retrieved candidates are used as semantic input in this case. (d) *V-COCO captions*: Since V-COCO is a subset of the COCO dataset [8], we use the corresponding captions as additional input along with the image. In the last row of the table, we also experiment with the oracle setting, where we assume we have access to the ground truth (GT) OA tuple. The strong performance of the oracle model indicates that refining the queries in this manner is an effective direction to guide the network to focus on more relevant candidates. The best performing approach in the non-oracle setting uses OA tuple. There is still a non-trivial gap between it and using the oracle, indicating that there is still room to improve the HOI detection accuracy by further improving the quality of the pre-generated OA tuple candidates.

It is interesting to note that using captions as additional input along with the image does not improve performance. This might be due to the fact that compared to the *object-action* candidate, the image captions can be noisy and sometimes too generic for the task (e.g. this photo has a horse etc.). Using only the $action$ approach achieves a slightly worse performance than using the OA tuple as expected, as the former has no information about the object category.

Number of Predictions as Semantics. We then ablate the different number of OA predictions candidate selected as the semantic inputs in Table 3c. We test with using topK, where K=1, 2, 4, 8 and 13 HOI predictions as semantics. We stop at K=13 because the maximum number of HOI ground truths for V-COCO in each image is 13. To better understand the results, we not only list the final mAP metric, but also add the precision and recall for the prediction in the table. Results show that K=4 gives the best performance, and the performance gradually decreases when moving to-

	mAP
Base (QPIC)	58.8
Base + Sem.	62.7
Base + Sem.+ Sp. (concat.)	62.9
Base + Sem.+ Sp. (multi.)	**63.7**

(a) *Support Feature Generator Module.*

	mAP
action only	63.1
OA	**63.7**
semantic retrieval	62.7
V-COCO captions	62.6
Oracle HOI GT	76.1

(b) *Semantic Inputs.*

#	Prec.	Rec.	mAP
1	85.3	23.4	62.9
2	75.1	41.2	63.3
4	62.5	68.6	**63.7**
8	37.1	81.5	63.4
13	25.1	89.6	62.8

(c) *Numbers of HOI predictions selected as semantics. Prec. is precision and Rec. is recall.*

	mAP
One-hot vector	63.0
GLOVE embedding [37]	63.1
CLIP text embedding [40]	**63.7**

(d) *Semantic embeddings.*

	mAP
Multi-variate parameter	62.3
Bi-variate parameter	62.0
Multi-variate spatial map	63.0
Bi-variate spatial map	**63.7**

(e) *Spatial feature designs.*

Table 3. *Design choices on the semantic and spatial features.*

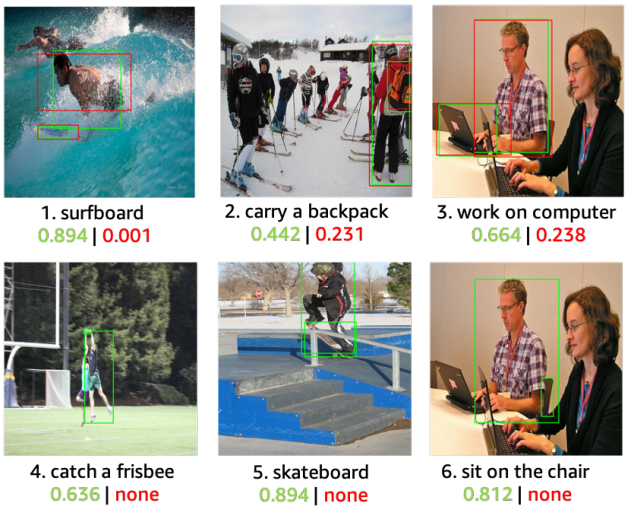


Figure 4. *Qualitative results of SSRT compared to QPIC. For each image, the detection outputs of SSRT are marked in green while the outputs of QPIC are marked in red. The prediction scores are presented in the captions. If no matched bounding box pairs are detected then the score is marked as none. We observe that SSRT improves over QPIC in mainly two aspects: (1) increasing the confidence scores of the interaction predictions (sample 1-3); and (2) successfully detecting the person, object and interactions that are completely missed in QPIC (sample 4-6).*

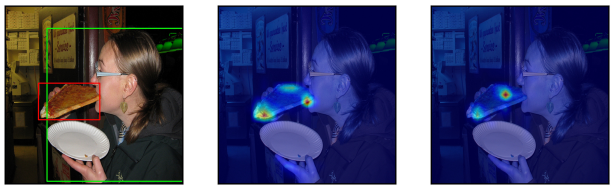
wards either direction (K=1 and K=13). As expected, K=1 gives the highest precision for prediction and K=13 gives the highest recall. But the optimal performance point (K=4) is in the middle, indicating that the trade-off between precision and recall of the prediction is important. Low recall corresponds to using lesser information for the query refiner block to produce representative enough queries for the decoder. Low precision affects the quality of input to the refiner block with increasing noise.

Semantic Embeddings. We then evaluate different embedding methods in Table 3d. We test with (1) a one-hot vector from the prediction; (2) the GLOVE [37] encoder, and (3) the CLIP [40] text encoder. Results show that all embeddings achieve good performance, while CLIP achieves the best. This may due to the fact that CLIP encoder is learnt from large-scale image text-pairs and hence generates a stronger semantic embedding for the HOI task than others. One-hot results also give good performance, indicating that using the pre-selected OA candidates itself can still provide guidance to refine the queries.

Spatial Feature Designs. We evaluate the performance of different spatial feature designs in Table 3e. For the relative spatial configuration (RSC) introduced in Sec. 3.2, we consider $(\Delta x_{oh}, \Delta y_{oh}, \Delta w_{oh}, \Delta h_{oh})$ to either follow a multi-variate distribution or follow two bi-variate distributions for $(\Delta x_{oh}, \Delta y_{oh})$ and $(\Delta w_{oh}, \Delta h_{oh})$. With each distribution, we explore two types of features: (1) only using parameters of distributions as features (Row 1 and 2). Specifically, for multi-variate distribution we use mean, variance and co-variance among all combinations of $\Delta x_{oh}, \Delta y_{oh}, \Delta w_{oh}, \Delta h_{oh}$ as the feature, and for bi-variate distribution we use the mean and variance of all $\Delta x_{oh}, \Delta y_{oh}, \Delta w_{oh}, \Delta h_{oh}$, plus only co-variance of $(\Delta x_{oh}, \Delta y_{oh})$ and of $(\Delta w_{oh}, \Delta h_{oh})$ as the feature. We concatenate them with semantic features as multiplication is not an option here; (2) we generate random samples from the distribution and then create the spatial map (Row 3 and 4) as introduced in Sec. 3.2. From Table 3e we can see that using spatial map always outperforms directly using parameters as features, which we believe is due to that spatial maps have a much higher dimension (2 x 64 x 64) than direct parameters (14 or 17) that can learn richer spatial configurations. In addition, the bi-variate distribution generated spatial map outperforms the multi-variate one.

Increased number of parameters of QPIC. Our design improves over QPIC by adding two novel modules (SFG and QR). To validate that SSRT's performance gain is from its design, rather than from its additional model capacity, we now experiment by increasing the number of parameter in QPIC's FFN to match those our approach (49.8M). Interestingly, QPIC's performance drops from 58.8 mAP to 57.9 mAP when its parameters are increased from 41.1M to 49.8M, likely due to overfitting. This clearly shows that

(a) Hold a tennis racket. Scores: SSRT: 0.856 | QPIC: 0.001.



(b) Eat a pizza. Scores: SSRT: 0.761 | QPIC: 0.512.

(c) Talk on the phone. Scores: SSRT: 0.612 | QPIC: none.

(d) Cut with knife. Scores: SSRT: 0.757 | QPIC: none.

Figure 5. *Visualization of the attention. We extract the attention map from the last layer of the decoder. In each sub-figure, from the left to the right are (1) the original image with the ground truth; (2) the attention map of our SSRT, and (3) the attention map of QPIC.*

SSRT performance (63.7 mAP) comes from our design.

Different ways of incorporating the information. Finally, we test three different ways of incorporating the semantic/ spatial information: (i) between the backbone and the encoder, (ii) as the input to the decoder, instead of using additional cross-attentions, (iii) to initialize the queries. None of these successfully matched SSRT's performance and didn't even improved over the performance of our QPIC baseline. This shows the importance of using such information properly. Upon analyzing these unsuccessful designs, we found that they were sensitive to the accuracy of the object-action (OA) selection and only when the ground truth OAs were used, their performance was better than QPIC. In contrast, SSRT is more robust to changes in OA selections, likely because it only uses the information as support features through additional cross-attention.

5.3. Qualitative Results

We show qualitative results of our SSRT and compare it with the baseline (QPIC). Fig. 4 shows results of examples selected from different interaction classes. We find that SSRT improves over QPIC mainly in two categories: (1) increasing the confidence scores of the action predictions (case 1-3); and (2) successfully detecting the person, object and actions that are completely missed (no bounding box output matches with GT) in QPIC (case 4-6). These improvement comes across different scenarios including: (1) small or nearly invisible objects (Sample 1, 4, 5, 6); (2) complex scenes (Sample 2); (3) multiple HOI predictions (Sample 3 and 6).

To further understand the network behavior, we compare the attention maps from SSRT and QPIC in Fig. 5. Specifically, we extract the visual attention maps of the query that predicts the marked person and object bounding boxes from the last layer of the decoder. In Fig. 5a, both QPIC and SSRT can localize the person and the object, but QPIC fails to predict the action with a high confidence while SSRT does. Looking at the attention map we can see the attention from QPIC is on the roughly correct region but very coarse and noisy, while it from SSRT is much more refined and focused on the area of the interaction (the hand). Similarly in Fig. 5b, SSRT achieves higher confidence than QPIC, as the attention is more refined and focused on the interaction area (the mouth and the hand), while QPIC just focuses on the pizza. For images in Fig. 5c and Fig. 5d, QPIC completely misses the prediction while SSRT detects the full correct HOI. We see from the attention map that SSRT is able to attend to the right area while QPIC fails. Overall we see that SSRT has more refined and sharper attention, and is able to focus on small objects in complex scenes.

6. Conclusion

We proposed SSRT, a one-stage semantic and spatial refined transformer for detecting HOIs. SSRT generates semantic and spatial features based on pre-selected human-object prediction candidates and leverages them to not only enrich the context but also guide the queries to attend to more related predictions. SSRT achieved SOTA performance on both V-COCO and HICO-DET datasets, demonstrating the effectiveness of our solution.

Limitation. Our approach requires fully-supervised HOI annotations for training, which are however extremely expensive to collect. In the future, it is important to explore novel HOI solutions that can learn from limited annotations and with less supervision.

Licenses. We use the following datasets: V-COCO (CC BY 4.0 license), HICO-DET [1] and code packages: QPIC [2], CLIP [3], GLOVE [4].

References

[1] http://www-personal.umich.edu/~ywchao/hico/. 8

[2] https://github.com/hitachi-rd-cv/qpic. 8

[3] https://github.com/openai/CLIP. 8

[4] https://github.com/stanfordnlp/GloVe. 8

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1, 2, 5

[6] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. 2, 5, 6

[7] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9004–9013, 2021. 1, 2, 6

[8] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 6

[9] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3550–3559, 2021. 2, 6

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[11] SM Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Koray Kavukcuoglu, and Geoffrey E Hinton. Attend, infer, repeat: Fast scene understanding with generative models. *NeurIPS*, 2016. 1

[12] Hao-Shu Fang, Yichen Xie, Dian Shao, and Cewu Lu. Dirv: Dense interaction region voting for end-to-end human-object interaction detection. In *The AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 1, 6

[13] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 6

[14] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *British Machine Vision Conference*, 2018. 1, 4, 5

[15] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2018. 4

[16] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 2, 5, 6

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 5

[18] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. Exploiting scene graphs for human-object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15984–15993, 2021. 1, 6

[19] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *European Conference on Computer Vision*, pages 584–600. Springer, 2020. 1, 2, 6

[20] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 495–504, 2021. 2, 6

[21] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14646–14655, 2021. 2, 6

[22] ASM Iftekhar, Satish Kumar, R Austin McEver, Suya You, and BS Manjunath. Gtnet: Guided transformer network for detecting human-object interactions. *arXiv preprint arXiv:2108.00596*, 2021. 2, 6

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021. 4

[24] Daisheng Jin, Xiao Ma, Chongzhi Zhang, Yizhuo Zhou, Jiashu Tao, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, Zhoujun Li, Xianglong Liu, et al. Towards overcoming false positives in visual relationship detection. *arXiv preprint arXiv:2012.12510*, 2020. 6

[25] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *European Conference on Computer Vision*, pages 498–514. Springer, 2020. 1, 6

[26] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 74–83, 2021. 1, 2, 6

[27] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. In *European Conference on Computer Vision*, pages 718–736. Springer, 2020. 1, 6

[28] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5

[29] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10166–10175, 2020. 2

[30] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu. Hoi analysis: Integrating and decomposing human-object interaction. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5011–5022. Curran Associates, Inc., 2020. 1, 2, 6

[31] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3585–3594, 2019. 1, 5

[32] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–490, 2020. 1, 2, 6

[33] Yang Liu, Qingchao Chen, and Andrew Zisserman. Amplifying key cues for human-object-interaction detection. In *European Conference on Computer Vision*, pages 248–265. Springer, 2020. 1, 2, 6

[34] Ye Liu, Junsong Yuan, and Chang Wen Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4235–4243, 2020. 1, 2

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 1

[36] I Loshchilov and Hutter Frank. Decoupled weight decay regularization. iclr 2019, 2017. 5

[37] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 5, 7

[38] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13018–13028, 2021. 6

[39] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–417, 2018. 1

[40] 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. 4, 5, 6, 7

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2

[42] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10419, 2021. 1, 2, 3, 5, 6

[43] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13617–13626, 2020. 1, 2, 3, 4, 5, 6

[44] Oytun Ulutan, Swati Rallapalli, Mudhakar Srivatsa, Carlos Torres, and BS Manjunath. Actor conditioned attention maps for video action detection. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 527–536, 2020. 2

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[46] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9469–9478, 2019. 1, 2

[47] Hai Wang, Wei-shi Zheng, and Ling Yingbiao. Contextual heterogeneous graph network for human-object interaction detection. In *European Conference on Computer Vision*, pages 248–264. Springer, 2020. 1, 6

[48] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. Deep contextual attention for human-object interaction detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5694–5702, 2019. 1

[49] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4116–4125, 2020. 1, 2, 6

[50] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *CVPR*, 2019. 1

[51] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, pages 418–434, 2018. 1

[52] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2, 3

[53] Dongming Yang, Yuexian Zou, Can Zhang, Meng Cao, and Jie Chen. Rr-net: Injecting interactive semantics in human-object interaction detection. *arXiv preprint arXiv:2104.15015*, 2021. 1

[54] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13319–13327, 2021. 1

[55] Yubo Zhang, Pavel Tokmakov, Martial Hebert, and Cordelia Schmid. A structured model for action detection. In *CVPR*, 2019. 1

[56] Xubin Zhong, Changxing Ding, Xian Qu, and Dacheng Tao. Polysemy deciphering network for human-object interaction detection. In *Proc. Eur. Conf. Comput. Vis*, 2020. 1, 6

[57] Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13234–13243, 2021. 2, 6

[58] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11825–11834, 2021. 1, 2, 6