PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

# Gtnet: guided transformer network for detecting human-object interactions

A. S. Iftekhar, Satish Kumar, R. Austin McEver, Suya You, B. Manjunath

A. S. M. Iftekhar, Satish Kumar, R. Austin McEver, Suya You, B. S. Manjunath, "Gtnet: guided transformer network for detecting human-object interactions," Proc. SPIE 12527, Pattern Recognition and Tracking XXXIV, 125270Q (13 June 2023); doi: 10.1117/12.2663936

SPIE.

Event: SPIE Defense + Commercial Sensing, 2023, Orlando, Florida, United States

# GTNet: Guided Transformer Network for Detecting Human-Object Interactions

A S M Iftekhar[a], Satish Kumar[a], R. Austin McEver[a], Suya You[b], and B.S. Manjunath[a]

[a]Department of Electrical & Computer Engineering, University of California, Santa Barbara
[b]US Army Research Laboratory

## ABSTRACT

The human-object interaction (HOI) detection task refers to localizing humans, localizing objects, and predicting the interactions between each human-object pair. HOI is considered one of the fundamental steps in truly understanding complex visual scenes. For detecting HOI, it is important to utilize relative spatial configurations and object semantics to find salient spatial regions of images that highlight the interactions between human object pairs. This issue is addressed by the novel self-attention based guided transformer network, GTNet. GTNet encodes this spatial contextual information in human and object visual features via self-attention while achieving state of the art results on both the V-COCO[1] and HICO-DET[2] datasets. Code is available online*.

**Keywords:** Activity detection, human-object interaction detection, scene graphs, scene understanding.

## 1. INTRODUCTION

Understanding human activity from images and videos is one of the main goals of computer vision based systems. One of the first steps in this process is the successful detection of human-object interactions (HOIs) in images, which consists of detecting the interactions between a human and an object while localizing their respective locations. In recent times, HOI has received much attention in the computer vision community due to its importance in tasks like action recognition in videos,[3–5] scene understanding,[6–8] and visual question answering.[9–11] Current research has integrated many different techniques ranging from attention mechanisms[12,13] to graph convolutional networks[14,15] to detect HOIs. In basic detection frameworks, human and object features are usually extracted with an object detection network. Then, interactions are predicted from the extracted features. To strengthen these features, one could use additional features such as relative spatial configurations,[12,13] human pose estimation,[16] or more accurate segmentation masks.[17]

However, to fully utilize the power of these additional features, the local spatial context needs to be leveraged more effectively. In computer vision tasks, context often refers directly to the background and surroundings of the objects or people of interest. In the following, we use "context" to refer to spatial regions localizing the human-object interactions. Few current works[13,15,18,19] have tried to find relevant spatial contexts by having separate attention mechanisms for humans and objects but do not leverage any additional features in the attention framework.

To this end, GTNet proposes a unique way to find spatial contextual information for detecting HOIs by leveraging spatial configurations (i.e. relative spatial layout between a human and an object) and object semantics (i.e. category of objects represented by word embeddings). This method has proven to be very effective, as can be seen in Figure 1, where our proposed Guidance Module improves performance over our model without the guidance mechanism.

Further author information: (Send correspondence to A. I.)
A.I.: E-mail: iftekhar@ucsb.edu,
S.K.: E-mail: satishkumar@ucsb.edu,
A.M.: E-mail: mcever@ucsb.edu,
S.Y.: E-mail: suya.you.civ@army.mil,
B.M.: E-mail: manj@ucsb.edu.
*https://github.com/UCSB-VRL/GTNet

Pattern Recognition and Tracking XXXIV, edited by Mohammad S. Alam, Vijayan K. Asari,
Proc. of SPIE Vol. 12527, 125270Q · © 2023 SPIE · 0277-786X · doi: 10.1117/12.2663936

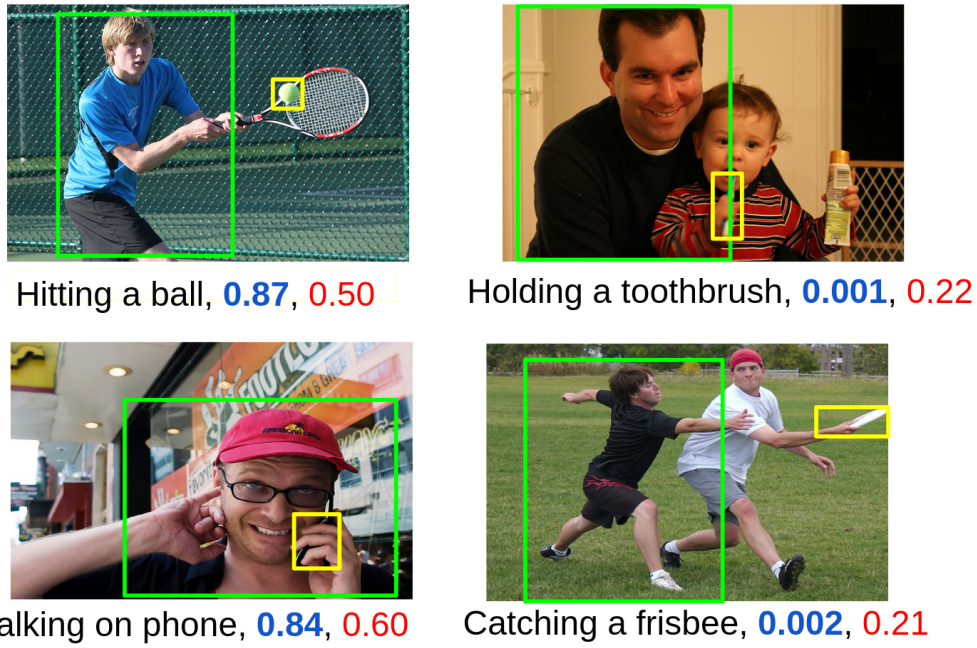


Figure 1. GTNet's performance with and without its guidance mechanism. Blue indicates GTNet's predictions with the guidance mechanism; red indicates without. Green bounding boxes indicate the human under consideration; yellow boxes indicate the object. Left column: Simple Scenarios (higher confidence score is better). Right column: Potentially Confusing Scenarios (lower confidence score is better). With the guidance mechanism, it is easier find salient spatial context for detecting different types of HOIs.

We take inspiration from Natural Language Processing (NLP) where the self attention[20] based Transformer architecture[21] has shown significant success in finding contextual information. To adapt the Transformer architecture for detecting HOIs, GTNet concatenates pairwise human and object features and uses them as queries to the attention mechanism. However, detection of HOIs often depend upon relative spatial configuration[12] and the type of the objects.[22] Therefore, we combine relative spatial configurations with object semantics to guide the queries throughout the network. With the help of our proposed guided attention, we successfully encode the spatial contextual information in these queries.

The proposed GTNet architecture can be seen in Figure 2. From the input image and the bounding boxes of the humans and the objects present in it, our baseline module (Section 3.1) extracts visual features. These types of visual features from a backbone network are being used across several domains[23–25] to capture useful information. We guide these visual features by pertinent spatial configurations and object semantics in our Guidance Module (Section 3.2). On top of these guided visual features, we develop the TX module, which enriches the visual features with relevant contextual information using attention (Section 3.3). Finally, we propose an early fusion strategy to make our final predictions (Section 3.4). We evaluated our network's performance on V-COCO[1] and HICO-DET[2] and achieve state of the art results on both of the datasets. Our contributions can be summarized as follows:

- We leverage pairwise spatial contextual information via a novel end to end guided self attention network for detecting HOIs. See Section 3.3.
- We design a guidance mechanism that combines relative spatial configurations and object semantics to guide our attention mechanism. See section 3.2.
- GTNet achieves state of the art results for the HOI detection task on both V-COCO and HICO-DET datasets. See section 4.3.

## 2. RELATED WORKS

**Human-Object Interaction:** With the introduction of benchmark datasets like V-COCO[1] and HICO-DET,[2] there is a plethora of works detecting human-object interactions.[12–19,22,26–31] Earlier works[32] in this area focus on the visual features of humans and objects. Many subsequent works[13,15,18] try to find spatial context for interactions on top of these visual features. Gao et al.[13] present a self-attention mechanism around individual humans and objects. T. Wang et al.[18] leverage this attention mechanism with a squeeze and excitation block[33] from object detection. Recently, graph-based architectures where humans and objects are considered nodes attempt to understand spatial context[15,19] for the structural relations. ConsNet[22] has leveraged word embeddings of the objects in this graph structure. Additionally, Hou et al.[34,35] have utilized object affordance to detect HOIs. A few recent works [36–39] have developed a one stage pipeline to detect HOIs rather than the two-stage (object detection + HOI detection) approach. We only compare our method with two stage HOI detection networks. However, none of the existing attention-based works try to utilize additional features like relative spatial configurations or object semantics to find richer spatial contextual features in the attention framework. In this aspect, GTNet proposes a pairwise attention network with a guidance mechanism for detecting HOIs by encoding spatial contextual information. We leverage spatial configurations and object semantic information to guide our attention network.

Many of the current works also use different additional features[2,12,16,17,27,40] ranging from pose information of humans to 3D representations of 2D images for detecting HOIs. Recently, VSGNet[12] utilizes relative spatial configurations to refine visual features. Inspired by this approach, we use spatial configurations to guide our self-attention mechanism. Further, we combine object semantics[22,41] with spatial configurations[2] for our guidance system. Our empirical results show that this combination performs better as a guidance mechanism.

**Transformer Network:** Recently, Transformer[21] based networks have achieved the state of the art performances in different vision tasks.[36,42,43] For detecting HOIs, one of the first attempts to use transformer like self-attention was.[44] Following that work, Girdher et al.[45] has developed the Actor Transformer network for detecting human activities in videos. Few recent works[36–39] have developed one stage Transformer networks for detecting HOIs. However, all these works rely only on self-attention mechanism to find salient context in the Transformer architecture. In contrast, we guide our joint feature representations of each human-object pair with spatial configurations and object semantics to find salient spatial context. Our state of the art performance over standard datasets validates our design choices for the self-attention network.

## 3. TECHNICAL APPROACH

In this section, we introduce our proposed GTNet architecture for Human-Object interaction (HOI) detection. Given an input image, the task is to generate bounding boxes for all humans and objects while detecting the interactions among them (e.g. a person hitting a ball). Each human-object pair can have multiple interactions. We use a pre-trained object detector to detect humans and objects in the images.

GTNet takes image features $\mathbf{F}$ as input. For extracting features, we use standard feature extractor networks (e.g., resnet[46] and efficientnet[47]). Consider a single image of size $c \times h \times w$ where $c, h$, and $w$ are the number of channels, height and width of the image. For a given image, $\mathbf{F}$ has a dimension of $C \times H \times W$ in the feature space. The relationship between $C, H, W$ and $c, h, w$ depends on the backbone network.

Along with image features, for each human and object our network takes bounding boxes $b_h$ and $b_o$ as inputs when there are $M \geq 1$ humans and $N \geq 1$ objects present in the image. As mentioned earlier, we use a pre-trained object detector to generate those bounding boxes. Our network will predict an interaction probability vector $\mathbf{p}_{HOI}$ for each human-object pair. In the next sections, we will describe different components of our network.

### 3.1 Baseline Module

Our baseline module extracts human and object feature vectors $\mathbf{f}_H$, and $\mathbf{f}_O$ from the input feature map $\mathbf{F}$. Following previous works[12,13] we use region of interest pooling followed by a residual block and average pooling to extract these feature vectors. Moreover, we use the overall feature map $\mathbf{F}$ to get generalized context information by extracting feature vector $\mathbf{f}_G$ with average pooling. Recent works like VSGNet[12] use this feature vector as a

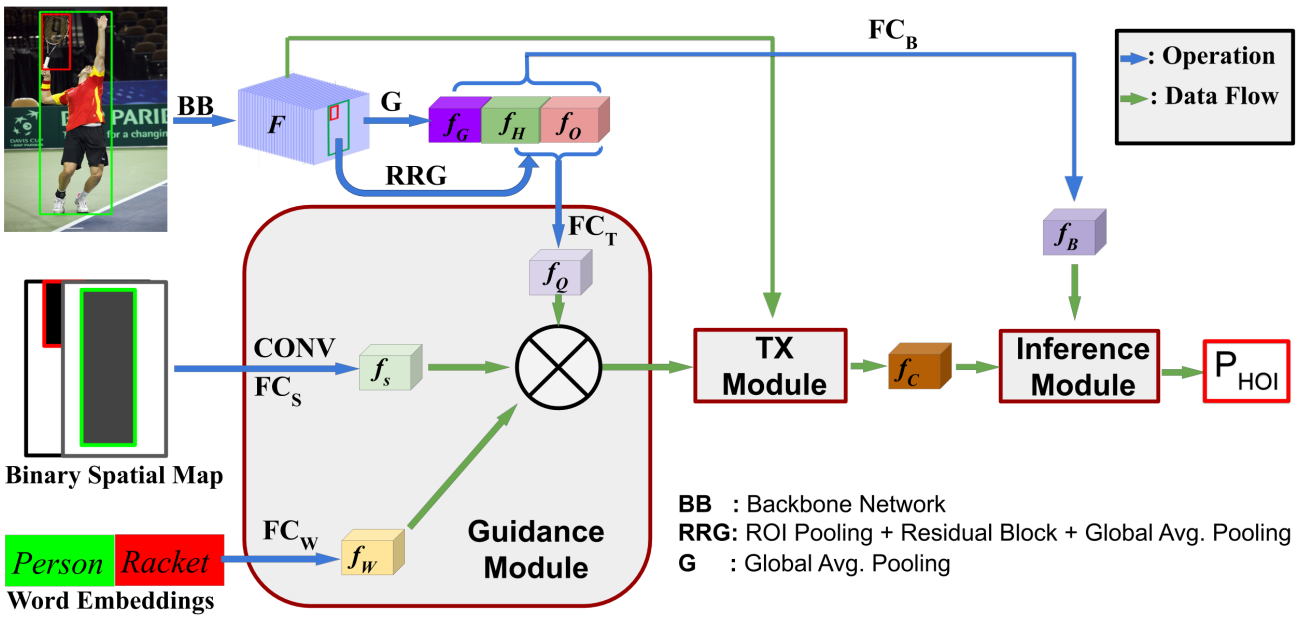


Figure 2. Model Overview. We extract human and object feature vectors from the input feature map, $\mathbf{F}$ via two RRG operations. For clarity we do not explicitly show two separate RRG operations. Human and object features are used to generate a query vector, $\mathbf{f}_Q$. Before feeding $\mathbf{f}_Q$ to the TX Module, we guide it via an element wise product with spatial ($\mathbf{f}_S$) and semantic ($\mathbf{f}_W$) guidance feature vectors. Inside the TX Module, contextual information is encoded to the guided query vector to generate a context-aware updated query vector $\mathbf{f}_C$. Finally, we make HOI predictions ($\mathbf{P}_{HOI}$) from the updated query and the baseline feature vectors in the Inference Module. Details of TX Module in Figure 3.

representation of context, we argue that it is not rich enough to account for interaction specific contexts. This is why we propose the guided attention mechanism to encode task specific spatial contextual information.

To get a joint representation of these feature vectors, we concatenate and project them in the same space using a fully connected (FC) layer.

$$\mathbf{f}_B = \mathbf{FC}_B(\mathbf{f}_H \parallel \mathbf{f}_O \parallel \mathbf{f}_G) \tag{1}$$

Here, $\parallel$ represents concatenation. This naive feature vector $\mathbf{f}_B$ is not adequate to detect all fine-grained HOIs. In the next sections, we describe how to refine and couple visual features with human-object pairwise spatial contextual information.

### 3.2 Guidance Module

Vaswani et al.[21] propose the Transformer network for processing sequential data in natural language processing (NLP). This network uses self attention[20] to find contextual dependence in sequential data. The original Transformer network introduces the concept of queries, keys, and values, which we adopt to the context of HOI detection. In this section, we explain the idea of queries and the mechanism to guide it. In the next section, we explain our attention mechanism.

Queries are defined as the pairwise joint representation of human and object feature vectors ($\mathbf{f}_H$, $\mathbf{f}_O$). We get the pairwise representation by concatenation and projection.

$$\mathbf{f}_Q = \mathbf{FC}_T(\mathbf{f}_H \parallel \mathbf{f}_O) \tag{2}$$

Here, $\mathbf{f}_Q$ is the query vector, and $\parallel$ represents concatenation operation. For a single human-object pair query vector has a length of $D$. This query vector will be used to find relevant spatial context in the feature map.

As mentioned in many previous works,[21,45] Transformer-like architectures do not perform well without positional information embedded in queries. For action recognition, Girdhar et al.[45] embedded bounding box sizes and locations in the queries. HOI detection is a more subtle task as, along with the size of the humans

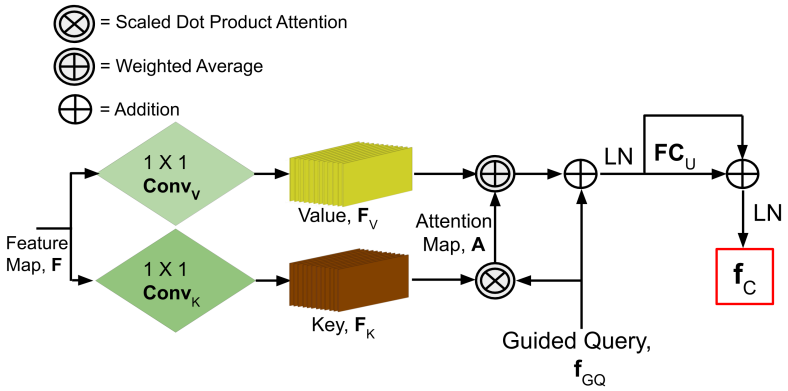


Figure 3. TX Module. From the input feature map, we generate key and value by $1 \times 1$ convolutions. With scaled dot-product attention, we produce an attention map for a particular human-object pair from the query and the key. This attention map is used to weigh the value to derive the contextually rich feature vector $\mathbf{f}_C$.

and the objects, relative configurations among them are also important. To this end, we propose our guidance module that uses relative spatial configurations and semantic representations of humans and objects to guide the attention mechanism.

**Spatial Guidance:** Relative spatial configurations have proven to be very useful[2,12,13,27] in detecting HOIs. We use a two-channel binary map with a dimension of $2 \times s \times s$ ($s$ is chosen as 64, see Table 5) to encode relative spatial configurations. For a human-object pair, the first channel is 1 in the location of the human-bounding box, whereas the second channel contains 1 in the location of the object bounding box. Everywhere else is zero in the binary map. Following,[2] we use two convolutional layers along with average pooling and linear projection($\mathbf{FC}_S$) to get a feature vector $\mathbf{f}_S$, representing the relative spatial configurations. We utilize $\mathbf{f}_S$ to guide the TX Module.

**Semantic Guidance:** Although $\mathbf{f}_S$ is a strong cue to detect HOIs, the information it conveys can be confusing without proper knowledge of the object. That is why we combine object semantics with spatial configurations in our guidance mechanism. Instead of just using a look-up table for identifying each object, we use word embeddings (vector representation of words) from the publicly available Glove[48] model. For our specific case, every detected object by the object detector is represented with a vector. For example, phone and football are presented by two different vectors. We express humans with one fixed vector. After concatenation of these human and object word embedding vectors together we get a combined feature vector $\mathbf{f}_W$ with linear projection ($\mathbf{FC}_W$). Along with $\mathbf{f}_S$, we utilize $\mathbf{f}_W$ in the guiding mechanism.

**Guidance Mechanism:** The information present in $\mathbf{f}_S$ and $\mathbf{f}_W$ is used to guide the queries sent to the TX Module. Guiding here refers to encoding relative spatial configurations and object semantics to the queries before feeding them to the TX Module. This can be achieved either by concatenation or by taking the element-wise product among the spatial, semantic ,and the query vectors. We use element-wise product as it is proven to be more effective guidance mechanism (See Table 3) than concatenation.

$$\mathbf{f}_{GQ} = \mathbf{f}_Q \circ \mathbf{f}_S \circ \mathbf{f}_W \tag{3}$$

Here, $\circ$ represents element wise dot product. We feed the guided query vector $\mathbf{f}_{GQ}$ to the TX Module.

### 3.3 TX Module

The TX module is our adaptation of the original Transformer architecture. Figure 3 shows the details of TX Module. This module encodes spatial contextual information to the guided query vectors via attention. We leverage the concept of keys and values from the original Transformer architecture in this mechanism.

Keys ($\mathbf{F}_K$) and values ($\mathbf{F}_V$) are generated from the input feature map ($\mathbf{F}$) by two separate $1 \times 1$ convolutions.

$$\mathbf{F}_K = \mathbf{conv}_K(\mathbf{F}) \tag{4}$$
$$\mathbf{F}_V = \mathbf{conv}_V(\mathbf{F}) \tag{5}$$

Here, $\mathbf{conv}_K$ and $\mathbf{conv}_V$ are two different $1 \times 1$ convolution operations. $\mathbf{F}_K$ and $\mathbf{F}_V$ can be thought of as two different representations of the input feature map. For a single image, both of them have the dimension of $D \times H \times W$.

The guided query vectors are used to search keys, $\mathbf{F}_K$ to find pairwise contextual information. Imagine a person is talking on phone in an image. Our guided query vector representing the person and the phone pair finds the important region (i.e. close to the ear) in $\mathbf{F}_K$ to detect the interaction: talking on phone.

This search process is done by scaled dot-product attention (equ. 6). In each spatial location of $\mathbf{F}_K$, we take a channel-wise scaled dot-product with the guided query vector. Softmax is applied to present the important context for a particular guided query in probabilities.

$$\mathbf{A} = \mathbf{Softmax}(\frac{\mathbf{f}_{GQ}\mathbf{F}_K^T}{\sqrt{\mathbf{D}}}) \tag{6}$$

For a particular guided query, $\mathbf{A}$ is the attention map where each element signifies the probability of that spatial location to be significant for detecting interactions. $\mathbf{A}$ has a dimension of $H \times W$ for each guided query vector. We use the attention map $\mathbf{A}$ to weigh $\mathbf{F}_V$ to get updated contextually rich query vectors. Following Transformer,[21] we use fully connected layers with residual connections in this process.

$$\mathbf{f}_C = \sum_{H,W} \mathbf{A} * \mathbf{F}_V \tag{7}$$

$$\mathbf{f}_C = LN(\mathbf{f}_C + \mathbf{f}_{GQ}) \tag{8}$$

$$\mathbf{f}_C = LN(\mathbf{f}_C + \mathbf{FC}_C(\mathbf{f}_C)) \tag{9}$$

Here, $LN$ means layer norm. It is used to stabilize the operations and prevent overfitting during training. Also, $*$ represents the Hadamard product between $\mathbf{A}$ and each channel of $\mathbf{F}_V$. The resultant weighted $\mathbf{F}_V$ was averaged over the spatial dimensions. $\mathbf{f}_C$ is the contextually enriched feature vector that will be used in the Inference Module.

We stack multiple TX Modules together to get better context representations in the updated queries like the original Transformer architecture.[21,45]

### 3.4 Inference Module

According to,[49] fusing features from different layers of a neural network increases the expressive capability of the features. Therefore, we fuse features from our Baseline Module with the TX Module. Moreover, we refine the naive visual features $\mathbf{f}_B$ in the same way as equ. 3.

$$\mathbf{f}_{BR} = \mathbf{f}_B \circ \mathbf{f}_S \circ \mathbf{f}_W \tag{10}$$

We concatenate $\mathbf{f}_B$, $\mathbf{f}_{BR}$, $\mathbf{f}_C$ together to increase the diversity in the final feature representation. With fully connected layers we make class-wise predictions for each human-object pair over this concatenated feature. Following,[12,27] we also generate an interaction proposal score, $\mathbf{b}_I$ using the baseline features for each human-object pair. This score represents the probability of interactions between a human-object pair irrespective of the class.

$$\mathbf{p}_I = \sigma(\mathbf{FC}_P(\mathbf{f}_B \parallel \mathbf{f}_{BR} \parallel \mathbf{f}_C)) \tag{11}$$

$$\mathbf{b}_I = \sigma(\mathbf{FC}_{P_B}(\mathbf{f}_B \parallel \mathbf{f}_{BR})) \tag{12}$$

Here, $\sigma$ represents a sigmoid non-linearity. For each human-object pair, we achieve our final predictions by multiplying these two individual predictions.

$$\mathbf{p}_{HOI} = \mathbf{p}_I \times \mathbf{b}_I \tag{13}$$

$\mathbf{p}_{HOI}$ has a length equal to the number of classes in considerations.

### 3.5 Loss Function

As HOI detection is a multi-label detection task (multiple interactions can happen with the same human-object pair), almost all prior works use binary cross entropy loss for each class to train the network. However, as pointed out by ,[12,15] confusing labels, missing labels, and mislabels are common in the HOI detection datasets. To handle those scenarios we utilize Symmetric Binary Cross Entropy (SCE) from [50] instead of only using binary cross entropy. This idea is derived from symmetric KL divergence and defined by:

$$\mathbf{SCE} = \alpha\mathbf{CE} + \beta\mathbf{RCE} = \alpha\mathbf{H}(p,q) + \beta\mathbf{H}(q,p) \tag{14}$$

where, **CE** is the traditional binary cross entropy, **RCE** is reverse binary cross entropy, **H** is entropy, **p** is target probability distribution and **q** is predicted probability distribution, $\alpha$ and $\beta$ are the weight values for each type of the loss. CE is useful for achieving good convergence, but it is intolerant to noisy labels. RCE is robust to noisy labels, as it compensates the penalty put on network by CE when the target distribution is mislabeled but the network is predicting a right distribution.For more details, please refer to .[50] We select $\alpha$ and $\beta$ as 0.5 to balance both kinds of losses.

## 4. EXPERIMENTS

In this section, we first describe our experimental setup and implementation details. We then evaluate GTNet's performance by comparing with previously state of the art methods. Finally, we validate our design choices via different ablation studies.

### 4.1 Experimental Setup

**Datasets:** There are two widely used publicly available datasets for the HOI detection task: V-COCO,[1] HICO-DET.[2]

V-COCO is a subset of the COCO dataset[51] and contains in total 10,346 images. The training, validation, and testing splits have 2,533; 2,867; and 4,946 images. Among its 29 classes, 4 do not contain any object annotations, and one of the classes has very few samples (21 images). Following previous works, we report our model's performance in the rest of the 24 classes. HICO-DET has in total 47,776 images: 38,118 for training, and 9,648 for testing. With 117 interaction classes, HICO-DET annotates in total 600 human-object interactions. Based on the number of training samples, the dataset is split into Full (all 600 HOI categories), Rare (HOI categories with sample number less than 10), and Non-Rare categories (HOI categories with sample number greater than 10).[2] We evaluate GTNet's performance in these categories.

**Evaluation Metrics:** We follow the protocol suggested by both datasets[1,2] and report our model's performance in terms of mean average precision (mAP). A prediction for a human-object pair is correct if the predicted interaction matches the ground truth and both the human and object bounding boxes have an intersection over union (IOU) score of 0.5 or higher with their respective ground truth boxes. For V-COCO, there are two protocols (Scenario 1 and Scenario 2) for reporting mAP.[1] When there is an interaction without any object (human only), in scenario 1 the prediction would be correct if the bounding box for the object is $[0, 0, 0, 0]$, in scenario 2 in these human only cases the bounding box for the object is not considered. For HICO-DET there are two settings: default and known. In default setting all images are considered to calculate AP for a certain HOI whereas in known setting only images that contain the particular object involved in that HOI are considered.

### 4.2 Implementation Details

As a backbone of GTNet, we experimented with different architectures and selected resnet-152[46] based on performance in the training set of V-COCO and HICO-DET. We also report our performance using resnet-50[46] for a fair comparison with existing methods. Output from the fourth residual block of resnet with a dimension of $1024 \times 25 \times 25$ was used as the the input feature map. By two separate $1 \times 1$ convolutions we generate key and value with a dimension of $512 \times 25 \times 25$.

For training, following the policy of previous works[12,15,22] we use human-object bounding boxes from a pre-trained Faster-RCNN[54] based object detector, the same as VSGNet .[12] For selecting the human and object

Table 1. Performance comparisons in the V-COCO[1] test set. Many current works do not report their models' performance in Scenario 2. The "Object Detector" column indicates on which dataset each method's object detector used for inference was trained. Ground Truth indicates that the method used ground truth bounding boxes for interaction predictions. Best results in each category are marked with **bold** and the second best results in those categories are marked with underline.

| Method | Object Detector | Feature Backbone | Scenario 1 | Scenario 2 |
|---|---|---|---|---|
| DRG[15] | COCO | ResNet50-FPN | 51.0 | - |
| VSGNet[12] | | ResNet - 152 | 51.8 | 57.0 |
| Wan et al.[16] | | ResNet50-FPN | 52.0 | - |
| Zhong et al.[28] | | ResNet - 152 | 52.6 | - |
| H. Wang et.al.[19] | | ResNet50-FPN | 52.7 | - |
| Kim et al.[29] | | ResNet - 152 | 53.0 | - |
| Liu et al.[17] | | ResNet-50 | 53.1 | - |
| ConsNet[22] | | ResNet-50 | 53.2 | - |
| IDN[31] | | ResNet-50 | 53.3 | 60.3 |
| OSGNet[52] | | ResNet-152 | 53.4 | - |
| Sun et al.[53] | | ResNet - 101 | 55.2 | - |
| GTNet (Ours) | | ResNet-50 | 56.2 | 60.1 |
| GTNet (Ours) | | ResNet-152 | **58.29** | **61.85** |
| VSGNet[12] | Ground Truth | ResNet-152 | 67.4 | 69.9 |
| GTNet (Ours) | | ResNet-50 | 71.45 | 73.50 |
| GTNet (Ours) | | ResNet-152 | **73.31** | **75.14** |

bounding boxes we use the same threshold (0.6 for human and 0.3 for object) as VSGNet. All projected feature vectors (baseline feature vector $\mathbf{f}_B$, query vector $\mathbf{f}_Q$, spatial guidance vector $\mathbf{f}_S$, semantic guidance vector $\mathbf{f}_W$, guided query vector $\mathbf{f}_{GQ}$, and context rich query vector $\mathbf{f}_C$) have a length of 512.

Hyper parameters of the network were selected by validating on V-COCO's validation set and a small split from the training set of HICO-DET. Our initial learning rate was 0.001 and the optimizer was Stochastic Gradient Descent (SGD) with a weight decay and a momentum of 0.9 and 0.0001. Our network was trained on GeForce RTX NVIDIA GPUs (one 2080 Ti for V-COCO, four 2080 Ti and four 1080 Ti for HICO-DET dataset) with a batch size of 8 per GPU. For each input image during training we randomly apply two augmentations from a set of augmentations (affine transformations, rotation, random cropping, random flipping, additive Gaussian noise, etc.). Our model has $\sim 60M$ parameters.

During inference, for each human-object pair, we multiply class-wise predictions $\mathbf{p}_{HOI}$ from equ. 13 with the detection confidence scores of the humans and objects. Also, we apply Low grade Instance Suppressive Function (LIS)[27] to improve the quality of the object detection confidence scores.

### 4.3 Results & Comparisons

In this section, we compare GTNet's performance with the current state of the art methods. As mentioned in Section 4.1, we follow the evaluation protocol suggested by the V-COCO[1] and HICO-DET[2] datasets. Moreover, for fair comparison we only compare our method with two stage HOI detection networks.

For testing on V-COCO, we use the object detection results from VSGNet,[12] which come from an object detector trained on COCO. For testing on HICO-DET, we use the object detection results provided by DRG,[15] which is used by the state of the art models.[12,15,31] These detections come from an object detector trained on COCO and finetuned on the training set of HICO-DET. GTNet achieves state of the art results in all datasets. Moreover, we use ground truth object detection results to isolate GTNet's performance from the effect of the object detector. Also, as most prior works either use resnet-152[12,28,29,52] or resnet-50[15,17,22,31] as feature backbone, we report our network's performance using both of these backbones.

In Table 1 and Table 2 GTNet's performance can be seen in V-COCO[1] and HICO-DET[2] datasets. Our network outperforms all other existing methods on the V-COCO dataset in both protocols on all object detectors. The current state of the art method[53] used a fusion method to fuse features from seven branches. We clearly outperform them with both of our backbones.

Table 2. Performance comparisons in the HICO-DET[2] test set. Def and ko mean default and known settings respectively. The "Object Detector" column indicates on which dataset each method's object detector used for inference was trained. Ground Truth indicates that the method used ground truth bounding boxes for interaction predictions. Best results in each category are marked with **bold** and the second best results in those categories are marked with underline.

| Method | Object Detector | Feature Backbone | Full(def) | Rare(def) | None-Rare(def) | Full(ko) | Rare(ko) | None-Rare(ko) |
|---|---|---|---|---|---|---|---|---|
| UnionDet[29] | COCO+HICO-DET | ResNet50-FPN | 17.58 | 11.72 | 19.33 | 19.76 | 14.68 | 21.27 |
| IPNet[55] | | Hourglass-104 | 19.56 | 12.79 | 21.58 | 22.05 | 15.77 | 23.92 |
| PPDM[56] | | Hourglass-104 | 21.1 | 14.46 | 23.09 | - | - | - |
| Bansal et al.[57] | | ResNet-101 | 21.96 | 16.43 | 23.62 | - | - | - |
| Hou et al.[30] | | ResNet-50 | 23.63 | 17.21 | 25.55 | 25.98 | 19.12 | 28.03 |
| ConsNet[22] | | ResNet-50 | 24.39 | 17.1 | 26.56 | - | - | - |
| DRG[15] | | ResNet50-FPN | 24.53 | 19.47 | 26.04 | 27.98 | 23.11 | 29.43 |
| IDN[31] | | ResNet-50 | 26.29 | 22.61 | 27.39 | 28.24 | 24.47 | 29.37 |
| VSGNet[12] | | ResNet-152 | 26.54 | 21.26 | 28.12 | - | - | - |
| ATL[34] | | ResNet-50 | 27.68 | 20.31 | 29.89 | 30.05 | 22.40 | 32.34 |
| FCL[35] | | ResNet-50 | 29.12 | **23.67** | 30.75 | 31.31 | **25.62** | 33.02 |
| GTNet (Ours) | | ResNet-50 | 26.78 | 21.02 | 28.50 | 28.80 | 22.19 | 30.77 |
| GTNet (Ours) | | ResNet-152 | **29.71** | 23.23 | **31.64** | **31.64** | 24.42 | **33.81** |
| iCAN[13] | Ground Truth | ResNet-50 | 33.38 | 21.43 | 36.95 | - | - | - |
| Li et al.[27] | | ResNet-50 | 34.26 | 22.9 | 37.65 | - | - | - |
| Peyre et al.[41] | | ResNet-50-FPN | 34.35 | 27.57 | 36.38 | - | - | - |
| VSGNet[12] | | ResNet-152 | 43.69 | 32.68 | 46.98 | - | - | - |
| IDN[31] | | ResNet-50 | 43.98 | **40.27** | 45.09 | - | - | - |
| GTNet (Ours) | | ResNet-50 | 44.71 | 33.80 | 47.97 | - | - | - |
| GTNet (Ours) | | ResNet-152 | **49.35** | 36.93 | **53.07** | - | - | - |

Moreover, GTNet outperforms all other existing methods with the HICO-DET object detector in the difficult default settings. Our much superior performance with ground truth object bounding boxes ($\sim$ 6 mAP improvement over the state of the art work) in non-rare category shows the effectiveness of our architecture. Also, though Transformer based network's needs good amount of data to train,[58] we manage to achieve second best performance in the RARE category. Actually, HICO-DET's rare category has 138 HOI classes with an average of 3 samples per class while 45 classes in rare category has only 1 sample in the training set. We expect to achieve much better performance in this category with a healthy number of samples.

The closest work to our network is DRG,[15] which utilizes a disjoint self-attention mechanism for each human and object in a dual relation graph network. By contrast, GTNet leverages the query, key, and value concept to encode pairwise contextual information in the queries along with a guiding mechanism and achieves significant improvement (7 mAP improvement in V-COCO and 5 mAP improvement in HICO-DET).

### 4.4 Ablation Studies

**Effect of various Components and Training Policy**: GTNet consists of several small modules and a unique training policy including the use of symmetric cross entropy loss, data augmentation, etc. In this section, we

Table 3. Ablation studies of GTNet in V-COCO test set. GTNet achieves state of the art performance with the guidance mechanism. It is interesting to observe that, in the same dataset, semantic guidance performs better than spatial guidance when tested independently. Also, it shows the effectiveness of symmetric cross entropy, interaction proposal score and data augmentation.

| | Scenario 1 | Scenario 2 |
|---|---|---|
| GTNet | **58.29** | **61.85** |
| without Spatial Guidance | 57.27 | 61.39 |
| without Semantic Guidance | 53.46 | 57.10 |
| without Guidance Module | 52.45 | 56.61 |
| without TX Module | 51.65 | 55.81 |
| without reverse cross entropy loss | 56.35 | 59.85 |
| without interaction proposal score | 57.27 | 60.97 |
| without data augmentation | 56.00 | 57.18 |
| guided by concatenation | 57.02 | 61.20 |

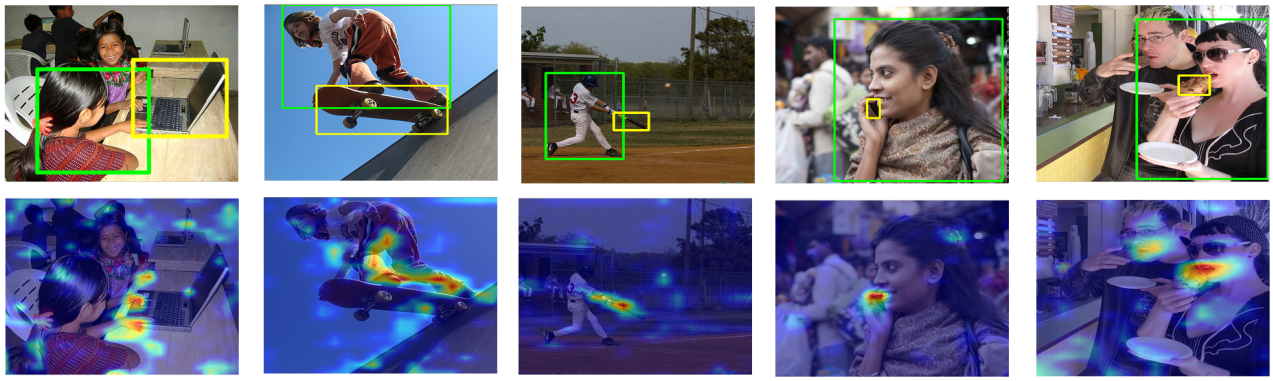


Figure 4. Qualitative Results on V-COCO test set. The top row is showing images with a particular human-object pair. The bottom row is showing class activation maps[59] generated for these pairs along with the interaction probability for particular interactions. The red region in the class activation map means the network is putting more attention in these areas. When there are same actions done by different human-object pairs (rightmost and leftmost images), the feature map gets activated in all relevant regions for the particular interaction. However, our pair wise guided attention strategy forces the network to consider the region significant to a particular pair.

examine their effectiveness in our overall performance in V-COCO test set (Table 3). As shown in Table 3 without the TX module, our baseline network achieves a mAP of 51.65. The performance improves slightly with the introduction of TX module. It is expected because, without the guidance mechanism, it is difficult to encode rich spatial contextual information to the visual features. With the help of the guidance mechanism, we improve our result by more than 6 mAP, highlighting the importance of the guiding mechanism in the attention framework. Also, our experiments demonstrate that semantic guidance is more effective than spatial guidance. Additionally, in Table 3 we have shown GTNet's performance without reverse binary cross entropy loss (just using binary cross entropy loss), data augmentation, and interaction proposal scores. All these actually helps our network to achieve superior performance. Moreover, as mentioned in Section 3.2 guidance can be achieved by either concatenation or product. As can be seen in Table 3, with guidance by concatenation we achieve 57.02 mAP compared to 58.29 mAP achieved with guidance by product.

**Number of Channels and Size of the Spatial Map** : We take $1 \times 1$ convolution over the input feature map to generate key and value with different number of channels. As can be seen in Table 4 with the increasing number of channels the network seems to perform a little better.

Similar trend can be seen in Table 5 with the increasing size of binary spatial map, $s$ in the spatial guidance mechanism. We choose 512 as the channel size for key and value and 64 as the size of the binary spatial map considering memory constraint in our GPU.

**Qualitative Results**: Figure 4 shows some qualitative results along with the class activation maps[59] for particular human-object pairs. As can be seen in Figure 4, GTNet correctly identifies the interactions with high confidence even when multiple interactions are happening together (image 1, 5). Moreover, from the class activation maps it is clear that our network is finding relevant context for the interacting human-object pair.

Table 4. Performance of GTNet with different number of channels for key and value in V-COCO test set.

| Number of Channels | Scenario 1 | Scenario 2 |
|---|---|---|
| 64 | 57.21 | 61.1 |
| 128 | 57.46 | 61.6 |
| 256 | 57.48 | 61.25 |
| 512 | **58.29** | **61.85** |

Table 5. Performance of GTNet with different size of binary spatial map, $s$ in V-COCO test set.

| $s$ | Scenario 1 | Scenario 2 |
|---|---|---|
| 4 | 56.84 | 60.63 |
| 16 | 57.30 | 60.99 |
| 32 | 57.95 | 60.63 |
| 64 | **58.29** | **61.85** |

## 5. DISCUSSION

### 5.1 Comparison with Other Attention Mechanisms

We compare GTNet with other recent attention mechanism based HOI detection networks: DRG,[15] and VSGNet.[12] DRG[15] considers pair-wise spatial-semantic features of humans and objects as nodes in their dual relation graph. With the use of self-attention, these features are aggregated to detect HOIs. However, we argue that using spatial-semantic features in the attention mechanism without the actual visual features will not help the network to retrieve the spatial context. In this respect, GTNet searches the whole feature map by specific queries, which are guided by spatial-semantics features to find relevant spatial context. Our superior quantitative results clearly validate our method. VSGNet[12] is the first attempt to utilize relative spatial configurations to refine visual features. However, VSGNet does not encode pair specific contextual information in the visual features as they take an average of the overall feature map as context. We utilize VSGNet's refinement strategy with object semantics to guide our attention mechanism.

### 5.2 Summary

We propose a novel guided self-attention network to leverage contextual information in detecting HOIs. Our pairwise attention mechanism utilizes a Transformer-like architecture to make visual features context-aware with the help of the guidance module. To the best of our knowledge, we are the first to propose a Guidance Module to guide the Transformer like attention mechanism to detect HOIs. We demonstrate by detailed experimentation that GTNet shows superior performance in detecting HOI and achieves the state of the art results in the standard datasets.

## 6. ACKNOWLEDGEMENT

This research is partially supported by the following grants: US Army Research Laboratory (ARL) under agreement number W911NF2020157; and by NSF award SI2-SSI #1664172. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of US Army Research Laboratory (ARL) or the U.S. Government.

## REFERENCES

[1] Gupta, S. and Malik, J., "Visual semantic role labeling," *arXiv preprint arXiv:1505.04474* (2015).

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

[3] Feichtenhofer, C., Fan, H., Malik, J., and He, K., "Slowfast networks for video recognition," in [*Proceedings of the IEEE international conference on computer vision*], 6202–6211 (2019).

[4] Li, C., Zhong, Q., Xie, D., and Pu, S., "Collaborative spatiotemporal feature learning for video action recognition," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 7872–7881 (2019).

[5] Ulutan, O., Rallapalli, S., Srivatsa, M., Torres, C., and Manjunath, B., "Actor conditioned attention maps for video action detection," in [*The IEEE Winter Conference on Applications of Computer Vision*], 527–536 (2020).

[6] Ren, H., El-Khamy, M., and Lee, J., "Deep robust single image depth estimation neural network using scene understanding.," in [*CVPR Workshops*], 37–45 (2019).

[7] Chen, P.-Y., Liu, A. H., Liu, Y.-C., and Wang, Y.-C. F., "Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 2624–2632 (2019).

[8] Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., and Gall, J., "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in [*Proceedings of the IEEE International Conference on Computer Vision*], 9297–9307 (2019).

[9] Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S., and Zhuang, Y., "Counterfactual samples synthesizing for robust visual question answering," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 10800–10809 (2020).

[10] Cadene, R., Ben-Younes, H., Cord, M., and Thome, N., "Murel: Multimodal relational reasoning for visual question answering," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 1989–1998 (2019).

[11] Li, L., Gan, Z., Cheng, Y., and Liu, J., "Relation-aware graph attention network for visual question answering," in [*Proceedings of the IEEE International Conference on Computer Vision*], 10313–10322 (2019).

[12] Ulutan, O., Iftekhar, A., and Manjunath, B. S., "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*], 13617–13626 (2020).

[13] Gao, C., Zou, Y., and Huang, J.-B., "ican: Instance-centric attention network for human-object interaction detection," in [*British Machine Vision Conference*], (2018).

[14] Qi, S., Wang, W., Jia, B., Shen, J., and Zhu, S.-C., "Learning human-object interactions by graph parsing neural networks," in [*Proceedings of the European Conference on Computer Vision (ECCV)*], 401–417 (2018).

[15] Gao, C., Xu, J., Zou, Y., and Huang, J.-B., "Drg: Dual relation graph for human-object interaction detection," in [*Proc. European Conference on Computer Vision (ECCV)*], (2020).

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

[17] Liu, Y., Chen, Q., and Zisserman, A., "Amplifying key cues for human-object-interaction detection," in [*European Conference on Computer Vision*], 248–265, Springer (2020).

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

[19] Wang, H., Zheng, W.-s., and Yingbiao, L., "Contextual heterogeneous graph network for human-object interaction detection," in [*European Conference on Computer Vision*], 248–264, Springer (2020).

[20] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y., "Show, attend and tell: Neural image caption generation with visual attention," in [*International conference on machine learning*], 2048–2057 (2015).

[21] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I., "Attention is all you need," in [*Advances in neural information processing systems*], 5998–6008 (2017).

[22] Liu, Y., Yuan, J., and Chen, C. W., "Consnet: Learning consistency graph for zero-shot human-object interaction detection," in [*Proceedings of the 28th ACM International Conference on Multimedia*], 4235–4243 (2020).

[23] Kumar, S., Iftekhar, A., Goebel, M., Bullock, T., MacLean, M. H., Miller, M. B., Santander, T., Giesbrecht, B., Grafton, S. T., and Manjunath, B., "Stressnet: Detecting stress in thermal videos," in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*], 999–1009 (2021).

[24] Wang, C., Pun, T., and Chanel, G., "A comparative survey of methods for remote heart rate detection from frontal face videos," *Frontiers in bioengineering and biotechnology* **6**, 33 (2018).

[25] Kumar, S., Torres, C., Ulutan, O., Ayasse, A., Roberts, D., and Manjunath, B., "Deep remote sensing methods for methane detection in overhead hyperspectral imagery," in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*], 1776–1785 (2020).

[26] Fang, H.-S., Xie, Y., Shao, D., and Lu, C., "Dirv: Dense interaction region voting for end-to-end human-object interaction detection," in [*The AAAI Conference on Artificial Intelligence (AAAI)*], (2021).

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

[28] Zhong, X., Ding, C., Qu, X., and Tao, D., "Polysemy deciphering network for human-object interaction detection," in [*Proc. Eur. Conf. Comput. Vis*], (2020).

[29] Kim, D.-J., Sun, X., Choi, J., Lin, S., and Kweon, I. S., "Detecting human-object interactions with action co-occurrence priors," in [*European Conference on Computer Vision*], 718–736, Springer (2020).

[30] Hou, Z., Peng, X., Qiao, Y., and Tao, D., "Visual compositional learning for human-object interaction detection," *arXiv preprint arXiv:2007.12407* (2020).

[31] Li, Y.-L., Liu, X., Wu, X., Li, Y., and Lu, C., "Hoi analysis: Integrating and decomposing human-object interaction," in [*Advances in Neural Information Processing Systems*], Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., eds., **33**, 5011–5022, Curran Associates, Inc. (2020).

[32] Gkioxari, G., Girshick, R., Dollár, P., and He, K., "Detecting and recognizing human-object interactions," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 8359–8367 (2018).

[33] Hu, J., Shen, L., and Sun, G., "Squeeze-and-excitation networks," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 7132–7141 (2018).

[34] Hou, Z., Yu, B., Qiao, Y., Peng, X., and Tao, D., "Affordance transfer learning for human-object interaction detection," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 495–504 (2021).

[35] Hou, Z., Yu, B., Qiao, Y., Peng, X., and Tao, D., "Detecting human-object interaction via fabricated compositional learning," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 14646–14655 (2021).

[36] Tamura, M., Ohashi, H., and Yoshinaga, T., "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*], 10410–10419 (2021).

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

[38] Kim, B., Lee, J., Kang, J., Kim, E.-S., and Kim, H. J., "Hotr: End-to-end human-object interaction detection with transformers," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 74–83 (2021).

[39] Zou, C., Wang, B., Hu, Y., Liu, J., Wu, Q., Zhao, Y., Li, B., Zhang, C., Zhang, C., Wei, Y., 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*], 11825–11834 (2021).

[40] Li, Y.-L., Liu, X., Lu, H., Wang, S., Liu, J., Li, J., and Lu, C., "Detailed 2d-3d joint representation for human-object interaction," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 10166–10175 (2020).

[41] Peyre, J., Laptev, I., Schmid, C., and Sivic, J., "Detecting unseen visual relations using analogies," in [*Proceedings of the IEEE International Conference on Computer Vision*], 1981–1990 (2019).

[42] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S., "End-to-end object detection with transformers," in [*European Conference on Computer Vision*], 213–229, Springer (2020).

[43] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929* (2020).

[44] Girdhar, R. and Ramanan, D., "Attentional pooling for action recognition," *arXiv preprint arXiv:1711.01467* (2017).

[45] Girdhar, R., Carreira, J., Doersch, C., and Zisserman, A., "Video action transformer network," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 244–253 (2019).

[46] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 770–778 (2016).

[47] Tan, M. and Le, Q. V., "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946* (2019).

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

[49] Yosinski, J., Clune, J., Bengio, Y., and Lipson, H., "How transferable are features in deep neural networks?," in [*Advances in neural information processing systems*], 3320–3328 (2014).

[50] Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., and Bailey, J., "Symmetric cross entropy for robust learning with noisy labels," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 322–330 (2019).

[51] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L., "Microsoft coco: Common objects in context," in [*European conference on computer vision*], 740–755, Springer (2014).

[52] Zhang, H.-B., Zhou, Y.-Z., Du, J.-X., Huang, J.-L., Lei, Q., and Yang, L., "Improved human-object interaction detection through skeleton-object relations," *Journal of Experimental & Theoretical Artificial Intelligence* , 1–12 (2020).

[53] Sun, X., Hu, X., Ren, T., and Wu, G., "Human object interaction detection via multi-level conditioned network," in [*Proceedings of the 2020 International Conference on Multimedia Retrieval*], 26–34 (2020).

[54] Ren, S., He, K., Girshick, R., and Sun, J., "Faster r-cnn: Towards real-time object detection with region proposal networks," in [*Advances in neural information processing systems*], 91–99 (2015).

[55] Wang, T., Yang, T., Danelljan, M., Khan, F. S., Zhang, X., and Sun, J., "Learning human-object interaction detection using interaction points," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 4116–4125 (2020).

[56] Liao, Y., Liu, S., Wang, F., Chen, Y., Qian, C., and Feng, J., "Ppdm: Parallel point detection and matching for real-time human-object interaction detection," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 482–490 (2020).

[57] Bansal, A., Rambhatla, S. S., Shrivastava, A., and Chellappa, R., "Detecting human-object interactions via functional generalization," in [*Proceedings of the AAAI Conference on Artificial Intelligence*], **34**(07), 10460–10469 (2020).

[58] Popel, M. and Bojar, O., "Training tips for the transformer model," *The Prague Bulletin of Mathematical Linguistics* **110**(1), 43–70 (2018).

[59] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A., "Learning deep features for discriminative localization," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 2921–2929 (2016).