Detecting Human-Object Relationships in Videos

Jingwei Ji Rishi Desai Juan Carlos Niebles
Stanford University
{jingweij, rdesai2, jniebles}@cs.stanford.edu

Abstract

We study a crucial problem in video analysis: human-object relationship detection. The majority of previous approaches are developed only for the static image scenario, without incorporating the temporal dynamics so vital to contextualizing human-object relationships. We propose a model with Intra- and Inter-Transformers, enabling joint spatial and temporal reasoning on multiple visual concepts of objects, relationships, and human poses. We find that applying attention mechanisms among features distributed spatio-temporally greatly improves our understanding of human-object relationships. Our method is validated on two datasets, Action Genome and CAD-120-EVAR, and achieves state-of-the-art performance on both of them.

1. Introduction

As we develop intelligent agents to understand images more comprehensively, the computer vision research problems we are solving have become more and more complex. The computer vision community has moved from classifying images and detecting objects, to detecting object relationships and understanding object interactions. In real-world applications, we often need to infer human behaviors from videos. In human-centered applications such as human-robot interaction, senior care [36], and healthcare [19], understanding the interactions people have with their environment is pivotal. One important problem at the heart of action recognition is detecting human-object relationships in videos: given the frames of a video, we would like to detect which objects a person is interacting with and classify the relationships between the person and objects.

Rather than generating scene graphs on static images [29, 53, 58, 45], human-object relationship (HOR) detection in videos focuses on the human and the *active* objects, that is the objects the person is actively interacting with. Unlike human-object interaction (HOI) detection [18, 7, 39, 33], HOR detection classifies not only the *verbs* that describe human actions, but also the *prepositions* between human and objects, such as "behind", "beneath" and "in". While many verbs are only associated with certain objects, prepositions are often applicable to numerous object categories.

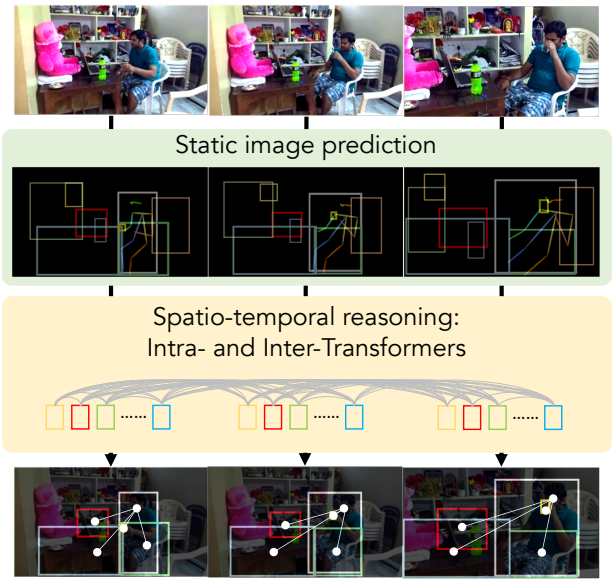

Figure 1: We tackle the problem of detecting human-object relationships in videos. Most prior approaches only model human-object relationships in images and perform static image predictions. We propose extra spatio-temporal reasoning on the top of static image predictions by intra- and inter-transformers.

Compared to scene graph generation and HOI detection on images, HOR detection in videos faces several challenges. First, the model needs to find out which objects are the *protagonists* of the scene. For instance, the clip in Figure 1 contains plenty of background objects that are not of our interest. How can we accurately focus on only the active objects? Second, the object detector, a key component in the detection model, will be confused during model training as video datasets [24, 60] typically only provide annotations on active objects. Without the knowledge of human-object interactions, a simple object detector will mistakenly fire on both the sat-on chair and those stacked chairs not relevant to the action. Third, videos can often be blurred at some

frames, and static image models have difficulty with performing inference on blurred frames. Considering these key issues, how can we leverage information from neighboring frames to produce more accurate predictions?

To tackle these issues, we propose a Human-Object Relationship Transformer (HORT) model for HOR detection in videos. Our model has two-stages: i) static image prediction and ii) using visual concepts from the first stage to recognize active objects and their relationships. HORT first performs static image prediction, where one can plug in an existing model from various choices of scene graph generation or HOI detection. In the second stage, HORT gathers the encoded visual concepts from the first stage (specifically the object, relationship, and human pose embeddings) and feeds them into transformers with intra- and inter-attention mechanisms. The attention mechanisms allow the model to integrate information from spatially and temporally scattered visual cues to find out which interactions are happening. In the transformer modules, we also pass messages from human pose and relationship features to object encoding, enabling an object scorer to focus on active objects.

To validate our HORT model, we have benchmarked its detection performance on two video datasets: Action Genome [24] and CAD-120-EVAR [60]. Our model outperforms state-of-the-art methods in scene graph generation and HOI detection. We have also conducted ablation studies to examine the contribution of each part of our model.

2. Related Work

Scene graph generation. Scene graphs are a symbolic representation of images, where objects are encoded as nodes and their relationships are encoded as connecting edges [26, 29]. This structured representation has bolstered many down-stream image tasks such as image retrieval [26, 41], visual question answering [25], visual reasoning [42], and image captioning [1]. A large body of work has focused on improving scene graph generation from single images. Lu *et al*. [35] propose to use both visual and language modules to generate scene graphs. Xu *et al*. [52] utilize RNNs to iteratively leverage node and edge information. Zellers *et al*. [57] highlight the regularly occurring graph structures existing in commonly used databases [29]. Li *et al*. [30] showcase the importance of context in hierarchical regions. Yang *et al*. [53] propose a relationship proposal network to prune edges in scene graphs and use attentional graph convolutional networks (GCNs) to integrate node information. Zhang *et al*. [58] introduce graphical contrastive losses. Guo *et al*. [16] apply a transformer on object features to explore contextual information among objects. Inspired by causal inference, Tang *et al*. [45] tackle the issue of biased representations. Zareian *et al*. [56] bridge commonsense knowledge graphs with scene graphs.

However, all these methods are limited to static images without modeling the spatio-temporal dynamics of relationships in videos. Moreover, most of the existing scene graph generation models implicitly assume a single-class relationship between each pair of objects [29], while this does not always hold, especially for human-object relationships [24] (e.g. `<person - looking at, holding, eating - food>` has three concurrent relationships).

Human-object interaction detection. Human-object interaction (HOI) detection [18, 7] aims to understand how a person is interacting with objects in an image. Our task, HOR detection, is similar to HOI detection but considers a broader class of edges. Relationships in HOI detection are *verbs*, such as "riding", "typing on", and "hugging", which often exclusively relate to certain object classes. Relationships in HOR detection can be *verbs* or *prepositions* [29, 24], such as "in", "behind", and "on the side of", which are more general and object-class-agnostic.

The task of HOI detection has resulted in a series of research [17, 39, 38, 48, 27, 33, 34, 23, 11]. Our model design shares the spirit of the multi-stream approach [7, 12, 50, 32], benefits from the knowledge of human poses [51, 31, 59], and leverages an attention mechanism among visual concepts. We further model the temporal dependencies between instances in our intra- and inter-transformer model, leading to better understanding of human-object interactions.

Transformer models in video analysis. Transformers [49] have emerged as one of the most powerful building blocks in natural language processing [9, 3]. Recent studies also showcase the capability of transformers on 2-D image tasks [37, 5, 2] and graph-structured data [55, 4]. Transformers have also been utilized in video analysis. Sun *et al*. [44] propose VideoBERT for action classification and video captioning on instructional videos. Girdhar *et al*. [15] introduce an action transformer network for action localization. Gavrilyuk *et al*. [14] tackle group activity recognition with actor transformers. Garcia *et al*. [13] include transformers in their model for video question answering. In our Human-Object Relationship Transformers, we leverage knowledge with both intra- and inter-attention from three visual concepts – human pose, object, and relationship – scattered in the 3-D spatio-temporal space.

3. Human-Object Relationship Transformers

The video HOR detection problem is defined as follows: we wish to build a model that takes a video clip as input, and on each frame outputs the location of the human, the locations of the active objects, and the multiple relationships between each human-object pair. Most recent scene graph generation [53, 58, 45] and HOI detection models [39, 32] for static images consist of three modules: a backbone image feature extractor, an object detection head, and another

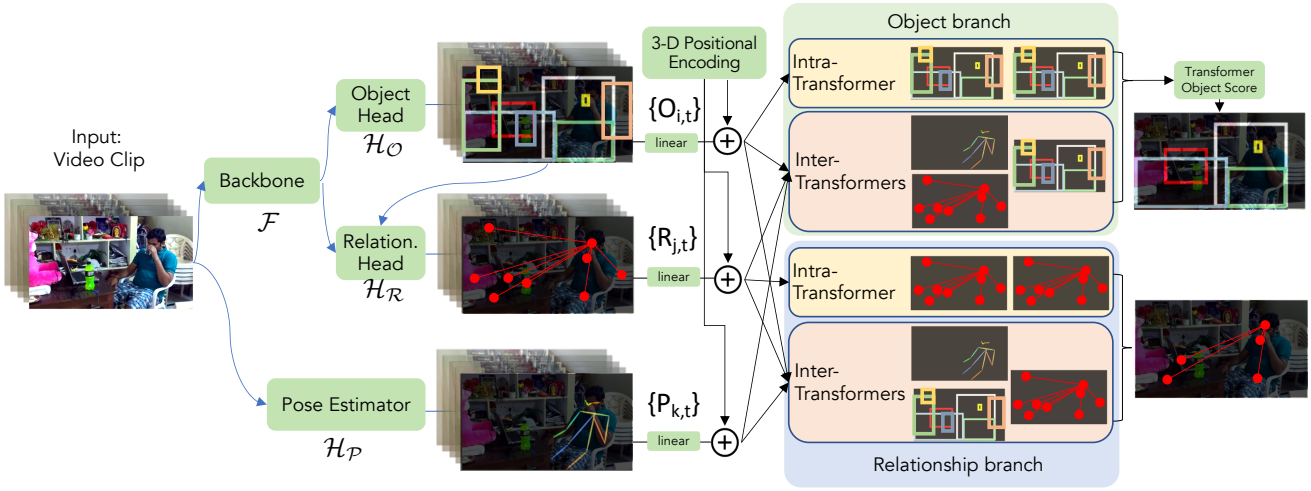

Figure 2: Our Human-Object Relationship Transformer (HORT) model. Our model extracts object, relationship and human pose features from the static image modules. Then these features are fed into an object branch and a relationship branch of intra- and inter-transformers. Finally the model generates human-object relationship detection by combining the object detection and relationship classification outputs.

head to predict the interactions or relationships between the detected objects. Our HORT model (Figure 2) adds an object branch and a relationship branch of intra- and inter-transformer modules on top of this framework. The transformer modules take in the temporal sequences of object, relationship, and pose features extracted from static images, integrate these features with an attention mechanism along both spatial and temporal dimensions, and finally generate more accurate human-object relationship detection.

3.1. Feature extraction on static images

Given a video clip $V = \{I_1, I_2, ..., I_T\}$, where I_t is the RGB frame at time step t, we first extract the feature map of each image: $x_t = \mathcal{F}(I_t)$. $\mathcal{F}$ refers to a backbone image feature extractor, typically implemented as a fully convolutional neural network. $x_t \in \mathbb{R}^{H' \times W' \times C}$ is the extracted image feature map that will be shared by both the object and relationship detection heads.

Our object detection head $\mathcal{H}_\mathcal{O}$ follows Faster R-CNN [40] and consists of a region proposal network and a box head. It takes in image feature maps x_t and generates bounding boxes $\{b_i^o\}_t$ of object proposals, each of which comes with an encoded object feature vector $O_i \in \mathbb{R}^{d_o}$ and a proposal confidence score $s_{i,static} \in (0, 1)$:

$$\{O_i\}_t, \{b_i^o\}_t, \{s_{i,static}^o\}_t = \mathcal{H}_\mathcal{O}(x_t), i \in \{1, ..., N_t\} \quad (1)$$

where N_t is the number of detected objects on frame t. N_t is typically larger than the number of ground truth objects in the scene, as the object detector outputs many *false positive* object proposals that are not being interacted with. The denotation $s_{i,static}^o$ indicates that these confidence scores are generated only with the knowledge of static images. $s_{i,static}^o$ highly depends on the appearance of objects rather than the context of human-object interactions.

The relationship head $\mathcal{H}_\mathcal{R}$ infers the relationships between each human-object pair detected by $\mathcal{H}_\mathcal{O}$. Following the designs of many relationship heads [53, 46, 58], we take the union box of each pair of human and object boxes b^o as a region of interaction b^r, pool an ROIAlign feature [20] from x_t, and then apply a neural network (e.g. ResNet-50 [21]) to extract a pairwise relationship feature $R_j \in \mathbb{R}^{d_r}$. $\mathcal{H}_\mathcal{O}$ also takes a set of object features $\{O_i\}_t$ as input and then outputs the logits $z_j \in \mathbb{R}^{C_R}$ (C_R referring to the number of relationship classes) for classifying the relationships between the j-th pair of human and object (in static image baseline methods). The function of $\mathcal{H}_\mathcal{R}$ is summarized as

$$\{R_j\}_t, \{b_j^r\}_t, \{z_j\}_t = \mathcal{H}_\mathcal{R}(x_t, \{b_i^o\}_t, \{O_i\}_t) \quad (2)$$

for all $j \in \{1, 2, ..., M_t\}$, where M_t denotes the number of human-object proposal pairs on frame t.

Research [54, 10, 31] has shown that understanding human-object interactions in static images can benefit from the knowledge of human poses. Hence, we believe that the temporal dynamics of human poses are helpful for inferring human-object relationships across time. Using a pose estimator $\mathcal{P}$, we generate human keypoints and determine bounding boxes of body parts following [10, 31]: $\{b_k^p\}_t = \mathcal{P}(I_t), k \in \{1, 2..., K_t\}$. K_t denotes the number of body parts in a person (head, shoulders, wrists, pelvic, knees, and ankles). Then we extract the feature $\{P_k\}_t \in \mathbb{R}^{d_p}$ for each body part from x_t: $\{P_k\}_t = \mathcal{H}_\mathcal{P}(x_t, \{b_k^p\}_t)$.

So far we have extracted three sets of static image features: objects $\{O_{i,t}\}$, relationships $\{R_{j,t}\}$, and human pose $\{P_{k,t}\}$. We will now show how to spatio-temporally integrate these features with our transformer models.

3.2. 3-D positional encoding

While Recurrent Neural Networks [22, 8] maintain the orders of tokens and Temporal Convolutional Neural Networks [47, 6] operate on temporal neighborhoods of features, the Transformer [49] is a sequential model that utilizes a fully-connected attention mechanism, thereby removing restrictions caused by positions and allowing dependencies between any pair of tokens to be modeled. However, in video analysis the knowledge of positions is still crucial. Before feeding the features extracted from static images into the transformer models, we need to reconstruct the positional information for each feature vector, i.e. where an object/relationship/pose feature is extracted from in the 3-D space of a video clip.

The original transformer model [49] uses a sinusoidal encoding for word positions. This positional encoding has been generalized to the x-y image plane in image applications [37, 5]. We further adapt this positional encoding for three dimensions, x, y, and time step t, so that the encoding indicates the position of each feature vector in 3-D space. After normalizing the x-y coordinates by $\tilde{x} = 2\pi x/W$, $\tilde{y} = 2\pi y/H$, the sinusoidal positional encoding along any dimension can be written as

$$\mathcal{PE}(*)_{2i} = \sin(*/10000^{2i/d_*}) \tag{3}$$

$$\mathcal{PE}(*)_{2i+1} = \cos(*/10000^{2i/d_*}) \tag{4}$$

where $*$ can either be $\tilde{x}$, $\tilde{y}$ or t. With the dimensionality of transformer input $d_{Tx} = 512$, we set $d_{\tilde{x}} = 128$, $d_{\tilde{y}} = 128$, $d_t = 256$, such that $d_{Tx} = d_{\tilde{x}} + d_{\tilde{y}} + d_t$. By concatenating the positional encoding from spatial and temporal dimensions, we get the 3-D positional encoding:

$$\mathcal{PE}(\tilde{x}, \tilde{y}, t) = \text{concat}(\mathcal{PE}(\tilde{x}), \mathcal{PE}(\tilde{y}), \mathcal{PE}(t)) \tag{5}$$

Note that the positional encodings are similar for features that are spatio-temporally close and different for those that are spatio-temporally far.

We use the (x, y) coordinates of the center point of each box b^o, b^r and b^p for its positional encoding. Because the dimensions of the features output by $\mathcal{H}_\mathcal{O}$, $\mathcal{H}_\mathcal{R}$ and $\mathcal{H}_\mathcal{P}$ may be different, we apply linear projections to align the dimensions into d_{Tx} for inputs to the transformer models. After adding the linearly projected features and positional encodings, we have the inputs for the transformers as follows:

$$O'_{i,t} = W_o^T O_{i,t} + \mathcal{PE}(b^o_{i,t}), W_o \in \mathbb{R}^{d_o \times d_{Tx}} \tag{6}$$

$$R'_{j,t} = W_r^T R_{j,t} + \mathcal{PE}(b^r_{j,t}), W_r \in \mathbb{R}^{d_r \times d_{Tx}} \tag{7}$$

$$P'_{k,t} = W_p^T P_{k,t} + \mathcal{PE}(b^p_{k,t}), W_p \in \mathbb{R}^{d_p \times d_{Tx}} \tag{8}$$

3.3. Intra- and Inter-Transformers

One of the key components of the original transformer model is computing the *multi-head self-attention* of representations, i.e. constructing an attention map between all pairs of features in a sequence, and then using this attention map to integrate features. Mathematically, the attention function computes the scaled inner product of a *query* feature sequence Q and a *key* feature sequence K to generate an attention map A. Then A is used to look up in a *value* feature sequence V. In the self-attention or intra-attention setup, Q, K, and V are linear projections of the *same* feature sequence. We are omitting the multi-head details here and we refer the readers to [49] or our supplementary materials for a more detailed description of the original intra-attention transformer model.

When detecting human-object relationships in videos, we need not only attention within each modality (object features $\{O'_{i,t}\}$ or relationship features $\{R'_{j,t}\}$), but also attention that is inter-modality. Pose and object features are critical cues for classifying relationships; pose and relationship features are helpful in determining which objects are being interacted with.

Our transformers are divided into two symmetric branches: an object branch and a relationship branch (Figure 3). For simplicity, we will only describe the details in the relationship branch (Figure 3(b)). The relationship branch consists of one intra-transformer and two inter-transformers, and each transformer contains an encoder and a decoder. The intra-transformer simply follows the original transformer design, where all Q, K, and V are linear projections of the relationship features R.

The first of the two inter-transformers in the relationship branch considers the attention between human poses and relationships. Intuitively, by simply looking at the human poses (especially a temporal sequence of them), one can make a reasonable impression of which relationships are occurring in the scene. Taking Figure 2 as an example: the pose suggests that the person is sitting somewhere and may be holding something close to his face.

The second inter-transformer utilizes attention among all three visual concepts: human poses, objects and relationships. The encoder computes inter-attention between the pose and object features, allowing the model to determine which objects are salient by looking at the trajectory of human poses and all objects. The encoder outputs a *pose-object memory* which is then passed into the decoder as a reference for classifying relationships.

We will now describe the architecture of inter-transformers. We prepare the input by flattening each set of features $\{O'_{i,t}\}$, $\{R'_{j,t}\}$ and $\{P'_{k,t}\}$ along spatial and temporal dimensions. Note that the positional information has been reserved in each feature vector, so flattening these features will not lose knowledge of proximity. We denote each resulting feature matrices as $O \in \mathbb{R}^{N \times d_{Tx}}$, $R \in \mathbb{R}^{M \times d_{Tx}}$, and $P \in \mathbb{R}^{K \times d_{Tx}}$, where $N = \sum_t N_t$, $M = \sum_t M_t$ and $K = \sum_t K_t$.

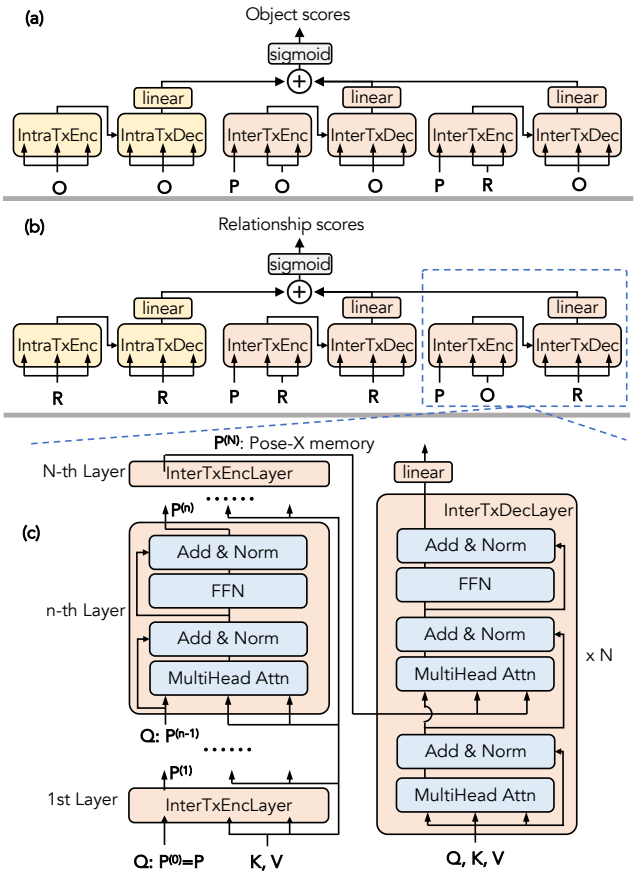

Figure 3: Architectures of the transformer models in HORT. We have (a) an object branch and (b) a relationship branch. Each branch contains one intra-transformer (IntraTx) and two inter-transformers (InterTx) with different choices of query, key, and value (input arrows from left to right) for the encoder and decoder. O, R and P stand for features of objects, relationships and poses, respectively. (c) The details of each encoder and decoder layer in an inter-transformer. The encoder takes pose features P as the initial query, and iteratively updates them with key and value features from objects or relationships. The encoder generates a Pose-X memory (X being object or relationship) as the key and value to the multi-head attention layers in the decoder.

The encoder consists of a stack of identical layers (with different weights). In the first inter-transformer, the n-th encoder layer takes R as the key and value and functions as follows:

$$P^{(n)} = \text{InterTxEncLayer}_{pr}^{(n)}(Q = P^{(n-1)}, K = V = R), \quad (9)$$

with the initial query $P^0 = P$. Similarly, the n-th encoder layer in the second inter-transformer takes O as the key and value, and encoder layers iteratively integrates object information into pose nodes:

$$P^{(n)} = \text{InterTxEncLayer}_{po}^{(n)}(Q = P^{(n-1)}, K = V = O). \quad (10)$$

The last encoder layer outputs a pose-relationship or pose-object memory matrix, depending on the encoder key and value. We use this memory matrix as the key and value in the multi-head attention layers in the decoder. The rest of the architecture of the inter-transformer follows the original transformer model [49].

In the relationship branch, each of the three transformers outputs a new relationship feature matrix. After computing the linear projections for each respective feature matrix, we add them together to create the logits for relationship classification. Prior work [30, 53, 58, 45] implicitly assumes only a single relationship can exist between each pair of subject and object; therefore, a softmax function is applied upon the logits to get the relationship scores. Because we are inferring multiple classes of relationships between human and objects (e.g. attentional, spatial, and contacting relationships [24]), we use the sigmoid function to generate per-class relationship scores $s^r \in (0,1)^{M \times C_R}$.

Having an architecture symmetric to the relationship branch, the object branch outputs object scores $s^o_{Tx} \in (0,1)^N$ for object proposals. These scores indicate if objects are salient in the context of human-object interactions.

3.4. Training and post-processing

Although the entire model can be trained end-to-end, we opt to pre-train the backbone and object detection head, fix their weights, and then only train the feature extractor in the relation head and the transformer models. This way, we can fairly compare our model with other baseline methods using the same backbone and object detector.

We utilize two loss functions while training our model: a binary cross entropy loss L_o for object saliency classification and a binary cross entropy loss L_r for relationship classification. We add the two losses for the total loss $L = L_o + \lambda L_r$.

Our method generates two scores for each object: s^o_{static} from the static image object detector and s^o_{Tx} from the object branch. We observed that s^o_{static} are highly biased towards the saliency of object appearance rather than whether an object is involved in any human-object interactions. False positive objects are often assigned with very high s^o_{static} scores. Conversely, s^o_{Tx} pay more attention to the interaction context as the object branch has integrated features from temporal sequences of poses, objects, and relationships; thus, s^o_{Tx} are usually low on false positive object proposals. We combine the two scores by choosing the minimum: $s^o = \min(s^o_{static}, s^o_{Tx})$. We have found that this fusion effectively suppresses false positive object proposals.

Finally, we compute a total score for each triplet $\langle$`subject - relationship - object`$\rangle$:

$$s = s^p * s^r * s^o, \quad (11)$$

where s^p is the confidence score for the human box generated in pose estimator. We rank all possible human-object

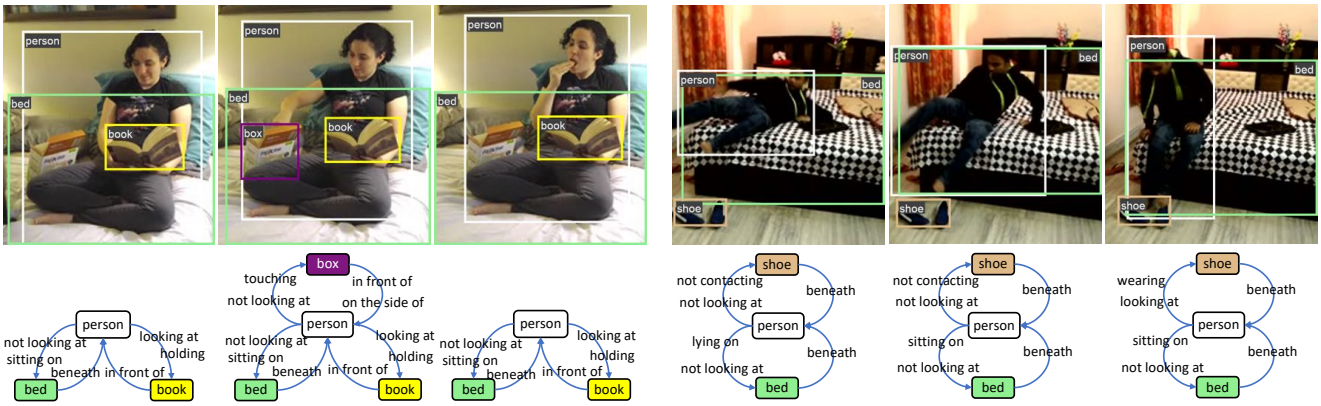

Figure 4: Examples of the detected human-object relationships in video clips in the Charades/Action Genome dataset. The detection formulates a multi-graph on each frame as we simultaneously predict multiple pairwise relationships. (Left) Note that the box is not detected as an active object when it is not interacted with. (Right) Relationships between the same human-object pair evolve with time.

relationship triplets in each frame by their total scores.

4. Experimental Results

4.1. Datasets

We have evaluated our HORT model over two third-person view video databases: Action Genome [24] and the re-annotated CAD-120-EVAR dataset [28, 60].

Action Genome. The Action Genome dataset [24] is built upon the crowdsourced videos from the Charades dataset [43], which captures indoor human activities and behaviors in daily life. Action Genome provides annotations of 476,229 bounding boxes of interacted objects and 1,715,568 relationship classes between the person and objects on 234,253 frames. Action Genome contains labels for 35 classes of objects and 25 classes of relationships. The relationships in Action Genome can be categorized into three types: *attentional* relationships indicating whether a person is looking at something, *spatial* relationships such as ⟨`chair - beneath - person`⟩, and *contacting* relationships indicating if a person is contacting an object and what type of contact is happening.

CAD-120-EVAR. The CAD-120 video dataset [28] consists of 4 subjects performing 10 different high-level household activities (e.g. arranging objects, taking food). Each subject performs each household activity 3 or 4 times, totaling 124 video sequences. In our experiments, we utilize the newly re-annotated version, which we call CAD-120-EVAR [60]. CAD-120-EVAR consists of 551 video clips with 32,327 frames. These frames are re-annotated to contain 6 classes of relationships between objects (e.g. holding, containing), the attributes of objects (e.g. open, closed), and the regions of interest of all of the objects in the frames.

4.2. Implementation details

The model is implemented in PyTorch. We use ResNet-101 [21] as our backbone image feature extractor and take the C4 features as x_t. For our experiments on both datasets, we pre-train the backbone and object detection head with the object detection task on Visual Genome [29]. For the Action Genome experiments, we further finetune the backbone and object head on Action Genome's training set. The same backbone and object detector is shared in all baseline experiments except for [33]. We do not finetune on CAD-120-EVAR as the "ground truth" object bounding boxes are generated by an object detector. We use an off-the-shelf Keypoint R-CNN [20] to estimate all human keypoints.

In our Action Genome experiments, we choose the clip length $T = 5$. Because Action Genome's annotation sampling rate is $\sim$1 FPS, our clip length effectively covers approximately 5 seconds on average. In our CAD-120-EVAR experiments, we set $T = 10$. We use a clip batch size of 4 in training on both datasets. Our models are trained on 4 Nvidia TITAN XP GPUs for 80,000 iterations with a learning rate starting at $5\text{e}{-4}$ and shrinking to $5\text{e}{-5}$ and $5\text{e}{-6}$ at iteration 30,000 and 50,000, respectively. For all the transformers, $d_{Tx} = 512$, 8 parallel heads are used, the feed forward dimension is 2048, and both the encoder and decoder contain 2 layers. Code will be released after acceptance.

4.3. HOR detection on Action Genome

Evaluation metrics. We follow the three standard evaluation modes for image-based scene graph prediction [35], which are also the evaluation metrics provided by Action Genome [24]: (1) predicate classification (PredCls) which assumes ground truth object classes and bounding boxes are given and only evaluates the predicate/relationship labels between each subject-object pair, (2) scene graph classification (SGCls) which assumes ground truth object bound-

Table 1: We compare our HORT model with recently proposed HOI detection models [39, 33] and image-based scene graph generation models [35, 52, 30, 53, 58, 46, 45]. Note that we use the same object detector for all baselines and our model except for PPDM [33]. AP_{50} stands for Average Precision at IoU threshold of 50%. @20 and @50 are abbreviation for recall@20 and recall@50. Our HORT model outperforms all baseline methods measured by all metrics.

Method	Object Detector		PredCls				SGCls				SGDet			
	Backbone	AP_{50}	image		video		image		video		image		video	
			@20	@50	@20	@50	@20	@50	@20	@50	@20	@50	@20	@50
GPNN [39]	ResNet-101	20.7	62.28	68.14	62.50	68.37	40.11	53.25	41.35	54.88	32.15	42.08	33.29	42.60
PPDM [33]	Hourglass-104	21.3	63.17	69.73	63.28	69.98	41.90	55.73	42.13	55.92	33.93	43.34	34.10	43.49
VRD [35]	ResNet-101	20.7	49.32	64.10	50.79	64.82	27.66	42.66	27.49	42.11	22.22	33.27	21.97	32.68
IMP [52]	ResNet-101	20.7	66.92	73.40	67.39	73.58	44.46	58.00	43.73	56.96	35.13	44.70	34.42	43.69
MSDN [30]	ResNet-101	20.7	67.22	73.43	67.73	73.60	44.72	58.20	44.12	57.21	35.27	44.79	34.65	43.81
Graph RCNN [53]	ResNet-101	20.7	67.31	73.60	67.84	73.80	45.02	58.46	44.49	57.46	35.53	45.05	34.95	44.09
RelDN [58]	ResNet-101	20.7	67.77	73.32	68.31	73.54	45.91	59.78	45.35	58.93	35.80	45.81	35.13	44.87
VCTree [46, 45]	ResNet-101	20.7	67.43	73.52	68.06	73.71	45.31	58.80	44.68	57.77	35.65	45.30	35.02	44.29
Temporal RelDN	ResNet-101	20.7	67.88	73.44	68.39	73.59	46.05	59.86	45.42	59.00	35.85	45.83	35.19	44.92
HORT (Ours)	ResNet-101	20.7	**71.67**	**76.16**	**72.39**	**76.66**	**47.68**	**62.56**	**47.11**	**61.61**	**37.19**	**47.76**	**36.51**	**46.67**

ing boxes are given and evaluates the triplet labels of ⟨subject - relationship - object⟩, and (3) scene graph detection (SGDet) which evaluates all predictions including bounding box locations and triplet labels. Action Genome further proposes the video version of these three metrics, where the per-frame measurements are first averaged in each video, then averaged across all videos in the test set. We report these metrics with recall@20 and recall@50, where recall@x computes the fraction of correct relationships in the top x ranked triplet predictions.

Baselines. We report the performance of various HOI detection and scene graph generation methods (Table 1). GPNN [39] is the only recent HOI detection method that also generalizes to video analysis. PPDM [33] is one of the state-of-the-art models in HOI detection. PPDM makes use of a different object detector than all other baselines. We pre-trained this hourglass-based object detector with the same curriculum as our Faster R-CNN detector: first on Visual Genome, then on Action Genome. PPDM's object detector achieves better performance (measured by AP_{50}), but PPDM does not perform as well as several other scene graph generation baselines in HOR detection.

Among the scene graph generation models, we compare with VRD [35], IMP [52], MSDN [30], Graph R-CNN [53], RelDN [58] and VCTree [46, 45]. When the backbone and object detector are fixed and shared by the baseline models, many models show similar performance, as they are designed for static images only. We have also extended the RelDN model with a simple approach of integrating temporal contextual information (Temporal RelDN in Table 1): when predicting relationships in both training and testing, the final logits of a frame is obtained by averaging the logits from a 5-frame temporal window around this frame.

Note that our reported measurements of baseline methods are significantly higher than those in [24]. This is because [24] has the restriction that only one relationship can be predicted between each pair of human and object during training and testing. Here, we remove this restriction, resulting in much higher measurements in *all baseline methods*. As a reference, we have also reported the evaluation comparisons with the single-relationship constraint in the supplementary material. Due to the intra- and inter-transformer models, HORT outperforms all baseline methods. Figure 4 illustrates examples of the predictions output by the HORT model. Please see the supplementary materials for more qualitative results.

4.4. Relationship classification on CAD-120-EVAR

CAD-120-EVAR does not provide manually labeled ground truth of object bounding boxes, so we do not train and test the object detector on CAD-120-EVAR. Therefore, the task of HOR detection is simplified to classifying the relationship between each pair of objects. The metric used in [60] is the accuracy of relationship classification, which is essentially the same as PredCls.

We report the classification accuracy of all 6 relationship categories in Table 2. Note that only *holding*, *not holding*, *contacting* and *apart* are human-object relationships, whereas *containing* and *separate* are actually relationships between *microwave* and other objects. Still, our model can handle the cases of non-human relationships as well. HORT outperforms both the baseline method reported in [60] and a RelDN [58] baseline we constructed.

4.5. Ablation study

We have conducted ablation experiments on Action Genome to inspect the effectiveness of each transformer, the object branch, and different positional encodings.

Transformer modules. In total, we have 6 transformers in our model: two branches for object scoring and relationship classification, each of which consists of an intra-transformer and two inter-transformers. We found that all of

Table 2: Accuracy of relationship classification on CAD-120-EVAR [28, 60]. Abbreviation for classes of relationships: HO - Holding; NH - Not holding; CTC - Contacting; AP - Apart; CTN - Containing; SP - Separate. Our method achieves similar or better performance on all relationship classes.

Method	HO	NH	CTC	AP	CTN	SP
`EVAR w/o obj` [60]	0.86	0.91	0.59	0.60	0.93	0.67
`EVAR w/ obj` [60]	0.82	**0.96**	0.80	0.96	0.95	0.96
`RelDN` [58]	0.88	**0.96**	0.88	0.95	0.95	0.94
`HORT (Ours)`	**0.89**	**0.96**	**0.98**	**0.97**	**0.96**	**0.97**

Table 3: Ablation study: effectiveness of different transformers. InterTx-2 refers to inter-transformers with two set of features as inputs, either pose-object in the object branch or pose-relationship in the relationship branch. InterTx-3 refers to inter-transformers with three modalities as inputs.

	IntraTx	InterTx-2	InterTx-3	SGDet-Img @20	SGDet-Img @50
Object branch	✓			36.03	46.10
	✓	✓		36.92	47.51
	✓	✓	✓	**37.19**	**47.76**
Relation branch	✓			35.93	45.96
	✓	✓		36.36	46.55
	✓	✓	✓	**37.19**	**47.76**

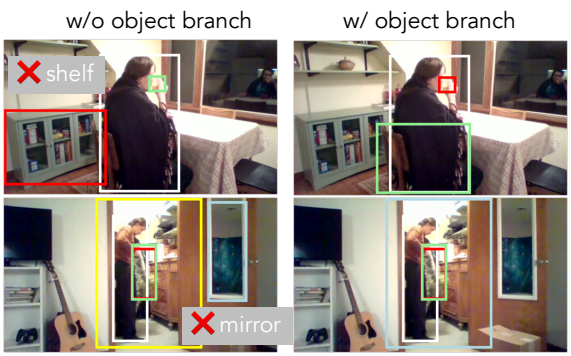

Figure 5: In these two pairs of frames, the single-frame object detector produces false positive proposals, such as a shelf and a mirror. With the transformer object scorer, our model re-ranks the object significance so proposals of inactive objects are removed.

the 6 transformers contribute to the human-object relationship detection performance. As shown in Table 3, adding each transformer into the model leads to a HOR detection performance gain measured by SGDet. The results do show a difference on the importance of the features. In both branch, the biggest performance gain comes from adding *pose-object* attention (InterTx-2 in the object branch and InterTx-3 in the relationship branch) compared to adding other types of attention. This phenomenon is intuitive as the temporal dynamics of human poses serve as strong cues for deciding which objects are being interacted with.

Transformer object scores. From Table 3, we can also tell that the transformer object scores s^o_{Tx} are crucial: if the entire object branch is ablated, the performance drops

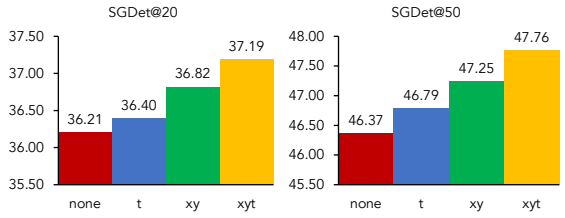

Figure 6: Ablation study: We measure the SGDet of our model with different types of positional encoding. t refers to temporal encoding only, xy to spatial encoding only, and xyt to the 3-D positional encoding we use in our full model.

to a level similar to static image baselines. We illustrate the function of s^o_{Tx} in Figure 5. Because the total object score is generated by taking the minimum of static image object detector scores s^o_{static} and transformer object scores s^o_{Tx}, false positive object proposals output by the object detector are assigned with lower confidence scores by the transformers, resulting in more precise object detection.

Positional Encoding. Positional encoding has been critical in many applications of transformer models including natural language processing [49, 9, 3], image recognition [37], and object detection [5]. For our spatio-temporal application, we apply positional encoding on three dimensions: x and y on the image plane and the time step t. As shown in Figure 6, adding 3-D positional encoding achieves superior HOR detection performance than the case with no positional encoding or only encoding either spatial or temporal positions. In the xy-only experiment, $d_x = d_y = 256$; in the t-only experiment, $d_t = 512$. Moreover, the data show that the positional encoding on the spatial dimensions are more important than the temporal counterpart.

5. Conclusion

In this paper, we propose a Human-Object Relationship Transformer (HORT) model for the problem of detecting human-object relationships in videos. Our model is composed of two stages, static image prediction and feature extraction, followed by intra- and inter-transformers performing spatio-temporal reasoning. By integrating features from different instances scattered across image planes and time steps, our model filters out false positive object proposals, identifies the active objects, and refines the relationship classification. We conducted experiments on two video datasets, Action Genome and CAD-120-EVAR, and showed how our model enables a better understanding of human-object relationships. We also inspected the contribution of each submodule in our ablation study, verifying the efficacy of our spatio-temporal reasoning.

Acknowledgment. We gratefully acknowledge the support of Panasonic and ONR (N00014-16-1-2127 and N00014-19-1-2477).

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016. 2

[2] Anonymous. An image is worth 16x16 words: Transformers for image recognition at scale. In *Submitted to International Conference on Learning Representations*, 2021. under review. 2

[3] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2, 8

[4] Deng Cai and Wai Lam. Graph transformer for graph-to-sequence learning. In *AAAI*, pages 7464–7471, 2020. 2

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020. 2, 4, 8

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 4

[7] 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. 1, 2

[8] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014. 4

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 8

[10] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 51–67, 2018. 3

[11] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *European Conference on Computer Vision*, pages 696–712. Springer, 2020. 2

[12] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018. 2

[13] Noa Garcia and Yuta Nakashima. Knowledge-based video question answering with unsupervised scene descriptions. In *Proceedings of the European Conference on Computer Vision*, 2020. 2

[14] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek. Actor-transformers for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 839–848, 2020. 2

[15] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019. 2

[16] Yuyu Guo, Jingkuan Song, Lianli Gao, and Heng Tao Shen. One-shot scene graph generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3090–3098, 2020. 2

[17] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1775–1789, 2009. 2

[18] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 1, 2

[19] Albert Haque, Arnold Milstein, and Li Fei-Fei. Illuminating the dark spaces of healthcare with ambient intelligence. *Nature*, 585(7824):193–202, 2020. 1

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3, 6

[21] 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. 3, 6

[22] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4

[23] 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. 2

[24] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020. 1, 2, 5, 6, 7

[25] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. *arXiv preprint arXiv:1705.03633*, 2017. 2

[26] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015. 2

[27] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. *arXiv preprint arXiv:2007.08728*, 2, 2020. 2

[28] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013. 6, 8

[29] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 1, 2, 6

[30] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1261–1270, 2017. 2, 5, 7

[31] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 382–391, 2020. 2, 3

[32] 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. 2

[33] 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, 7

[34] 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. 2

[35] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer, 2016. 2, 6, 7

[36] Zelun Luo, Jun-Ting Hsieh, Niranjan Balachandar, Serena Yeung, Guido Pusiol, Jay Luxenberg, Grace Li, Li-Jia Li, N. Downing, Arnold Milstein, and Li Fei-Fei. Computer vision-based descriptive analytics of seniors' daily activities for long-term health monitoring. In *Proceedings of the 3rd Machine Learning for Healthcare Conference*, volume 85, pages 1–18, 2018. 1

[37] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018. 2, 4, 8

[38] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting rare visual relations using analogies. 2019. 2

[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, 2, 7

[40] 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. 3

[41] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015. 2

[42] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8376–8384, 2019. 2

[43] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 6

[44] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7464–7473, 2019. 2

[45] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3716–3725, 2020. 1, 2, 5, 7

[46] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6619–6628, 2019. 3, 7

[47] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497. IEEE, 2015. 4

[48] 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. 2

[49] 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, 4, 5, 8

[50] 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. 2

[51] 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 Conference on Computer Vision and Pattern Recognition*, 2019. 2

[52] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2017. 2, 7

[53] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. *arXiv preprint arXiv:1808.00191*, 2018. 1, 2, 3, 5, 7

[54] Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 17–24. IEEE, 2010. 3

[55] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. Graph transformer networks. In *Advances in Neural Information Processing Systems*, pages 11983–11993, 2019. 2

[56] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Bridging knowledge graphs to generate scene graphs. In *European Conference on Computer Vision*, pages 606–623. Springer, 2020. 2

[57] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. *arXiv preprint arXiv:1711.06640*, 2017. 2

[58] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11535–11543, 2019. 1, 2, 3, 5, 7, 8

[59] Xubin Zhong, Changxing Ding, Xian Qu, and Dacheng Tao. Polysemy deciphering network for human-object interaction detection. In *Proc. Eur. Conf. Comput. Vis*, 2020. 2

[60] Tao Zhuo, Zhiyong Cheng, Peng Zhang, Yongkang Wong, and Mohan Kankanhalli. Explainable video action reasoning via prior knowledge and state transitions. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 521–529. ACM, 2019. 1, 2, 6, 7, 8