LIGHTEN: Learning Interactions with Graph and Hierarchical TEmporal Networks for HOI in videos

Sai Praneeth Reddy Sunkesula
Indian Institute of Technology Bombay
Mumbai, Maharashtra, India
praneeth20@cse.iitb.ac.in

Rishabh Dabral
Indian Institute of Technology Bombay
Mumbai, Maharashtra, India
rdabral@cse.iitb.ac.in

Ganesh Ramakrishnan
Indian Institute of Technology Bombay
Mumbai, Maharashtra, India
ganesh@cse.iitb.ac.in

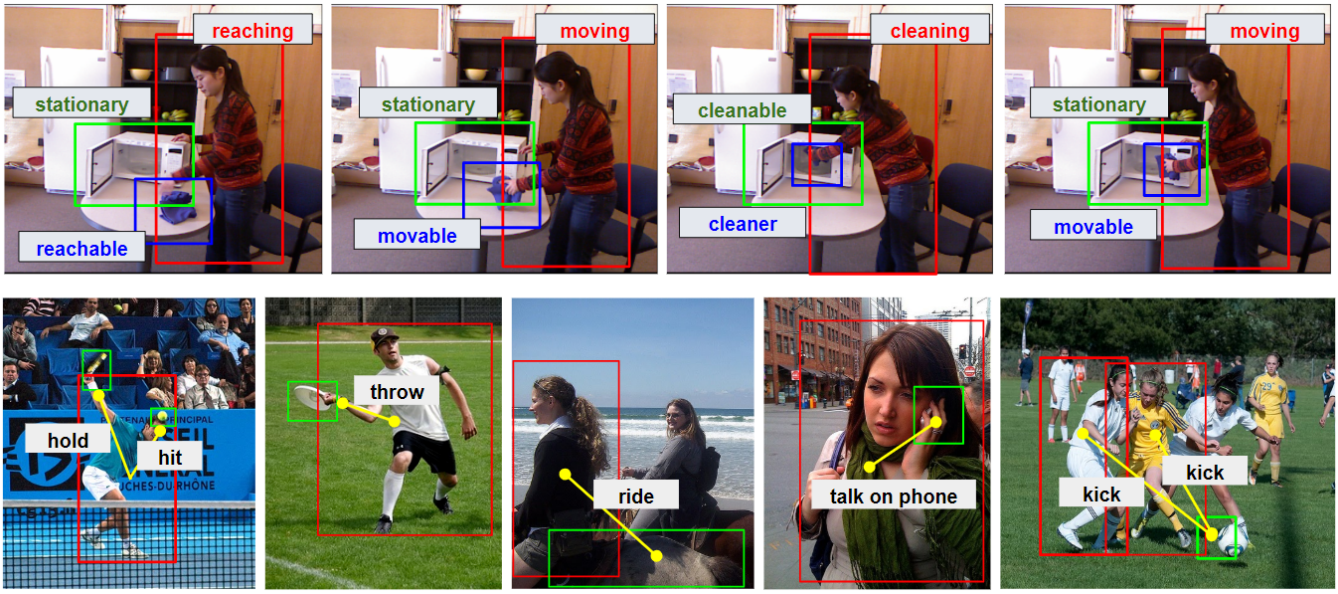

Figure 1: Illustration of human-object interaction detection in video (CAD-120) and image (V-COCO) settings

ABSTRACT

Analyzing the interactions between humans and objects from a video includes identification of the relationships between humans and the objects present in the video. It can be thought of as a specialized version of Visual Relationship Detection, wherein one of the objects must be a human. While traditional methods formulate the problem as inference on a sequence of video segments, we present a hierarchical approach, LIGHTEN, to learn visual features to effectively capture spatio-temporal cues at multiple granularities in a video. Unlike current approaches, LIGHTEN avoids using ground truth data like depth maps or 3D human pose, thus increasing generalization across non-RGBD datasets as well. Furthermore, we achieve the same using only the visual features, instead of the commonly used hand-crafted spatial features. We achieve state-of-the-art results in human-object interaction detection (88.9% and 92.6%) and anticipation tasks of CAD-120 and competitive results on image based HOI detection in V-COCO dataset, setting a new benchmark for visual features based approaches. Code for LIGHTEN is available at https://github.com/praneeth11009/LIGHTEN-Learning-Interactions-with-Graphs-and-Hierarchical-TEmporal-Networks-for-HOI

KEYWORDS

Human-Object Interaction, Visual Relationships, Hierarchical RNN, Spatio-Temporal Graph Modelling

ACM Reference Format:
Sai Praneeth Reddy Sunkesula, Rishabh Dabral, and Ganesh Ramakrishnan. 2020. LIGHTEN: Learning Interactions with Graph and Hierarchical TEmporal Networks for HOI in videos. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413778

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413778

1 INTRODUCTION

A key element of Scene Understanding is perception and interpretation of humans and the associated interactions. While human

perception typically involves inferring the physical attributes about the humans (detection [5, 35, 43, 50], poses [3, 4, 8, 25, 28, 41], shape [13, 20, 29, 30], gaze [44] etc.), interpreting humans involves reasoning about the finer details relating to human activity [6, 24, 27, 48, 49], behaviour [26, 34], human-object visual relationship detection [23, 33, 36, 37, 39, 40], and human-object interactions [23, 32, 33, 36, 37, 39, 40, 42]. In this work, we investigate the problem of identifying Human-Object Interactions in videos. Given a video stream, the goal is to identify the objects interacting with the humans while also estimating the kind of interaction, *eg*., holding the cup, placing the bowl, moving the furniture, *etc.* The availability of such information can be crucial in understanding the finer details of human behaviour than in, say, action recognition. Such information has the potential to facilitate downstream applications like unmanned supermarkets, surgery documentation, robotics, *etc.*

While investigating into the HOI problem, we especially focus on video settings. There has been a significant amount of research on HOI with images [21, 32, 42, 46], thanks to the availability of V-COCO [9] and HICO [1] datasets. However, learning human-object interactions within videos is challenging and relatively less explored owing to multiple reasons. *Firstly*, it requires the model to account for the changing orientations of objects in the scene with respect to the humans. This makes it difficult to extend the image-based approaches that use the RoI features of the union of human and object to the video setting. *Secondly*, the unavailability of large scale video datasets (except CAD-120 [15]) makes it difficult to train an HOI model that is generic, and performs well for in-the-wild videos. *Finally*, the interaction definitions tend to become confusing when defined for a video, *e.g*., *placing* vs. *moving* vs. *reaching*, *opening* a jar vs. *closing* a jar, etc. In spite of these challenges, videos allow for exploiting temporal visual cues that are, otherwise, absent in images.

Most existing methods are designed to work in either the image setting[21, 42], or the video setting [12, 16] but not both. Recently, Qi et al. [32] proposed a graph-parsing based method that fits into to both the settings. While the method indeed achieves state-of-the art results in video setting, it does so by using carefully designed and pre-computed hand-crafted features such as SIFT [31] transforms, object centroids, 3D poses, object depths, etc. which were originally proposed in [15]. It is worth noting that these features were derived from the ground-truth data provided by the CAD-120 dataset. It is straightforward to see that using ground-truth based features for estimating HOI would not allow the method to perform well on in-the-wild videos because such features may either not be available (3D pose) or may be noisy and inconsistent across frames (object bounding boxes, centroids, *etc.*)

With these caveats in mind, we propose a hybrid approach that uses Graph Convolutional Network (GCN) and hierarchical RNNs, LIGHTEN, for detecting human-object interactions from videos that *does not* rely on hand-crafted features. We use pure visual features derived from a re-trainable off-the-shelf network to represent the inputs to LIGHTEN and demonstrate strong performance on the CAD-120 dataset. Furthermore, The proposed network is designed to leverage the spatio-temporal cues that are crucial to disambiguate confusing interactions. Specifically, we design a two-level architecture which, i) performs graph-based spatial embedding extraction from the video and learns temporal reasoning functions at the frame level, followed by ii) a segment level temporal network which learns inter-segment temporal cues from previous segments, for regressing the human sub-activities and object affordances. The temporal functions are designed to learn the temporal relationships between human-object pairs across the video.

Despite not using the ground truth based pre-computed features and in spite of the small amount of data available for training from videos, our visual input based model achieves state-of-the-art performance on sub-activity, affordance detection tasks, setting a strong baseline for the future of such methods. When used with the segment level pre-computed features, the segment-level temporal model of our proposal performs at par with the state-of-the-art methods. Finally, despite being designed for video-based tasks, our method also demonstrates competitive performance on the V-COCO dataset that corresponds to the image setting.

In summary, we make three contributions in this paper in the form of our model, LIGHTEN: *First*, we propose a generalizable, multi-level method for identifying Human-Object Interactions from videos. To the best of our knowledge, ours is the first that performs video-based HOI estimation purely from learnt visual features. *Second*, we setup a new baseline for such methods as ours on CAD-120 dataset while also approaching competitive results with methods that are either purely image-based or purely video-based. *Third*, we show how LIGHTEN naturally lends itself to static, image-based settings.

2 RELATED WORK

Human-Object Interaction detection has been a well researched problem. In this section, we discuss the existing literature from two broad viewpoints: static (images) and dynamic (videos).

HOI from images: Prior to deep learning, initial works on HOI from images were based on using hand-crafted features such as SIFT, HOG, etc. Among such works, Yao *et. al.* [48] learned the bases of actions and parts to reason about HOI. Likewise, Hu *et. al.* [11] used HOI exemplars to model the spatial relationships between the human and the objects. A problem like Human-Object Interaction should be amenable to the use of structure based reasoning, by virtue of the fact that HOI requires detection of humans and objects and their spatial interactions that are expected to persist temporally. Toward this, Yao *et. al.* [47] define *grouplets* as a feature encoder for capturing structural information, Delaitre *et. al.* [6] construct structure-aware feature representations that are trainable with an SVM.

Recently, deep learning based methods, bolstered by the availability of large amounts of in-the-wild training data [1, 9] have lead to significantly improved performance in HOI detection. Among such methods, Li *et. al.* [21] proposed to learn the knowledge about the *interactiveness* between the humans and object categories from HOI datasets and use this knowledge as a prior while performing HOI detection. For understanding the interactions, it has also been argued that human pose provides useful cues about the type of interaction. For example, a human *opening* a jar will have a significantly different pose than when the human is *reaching* for a jar. Several methods have attempted to leverage the human pose information in their pipelines. Wan *et. al.* [42] propose a pose-aware network

architecture that employs a multi-level feature strategy, thereby dealing with the problem at three levels of granularity: overall interactions (covering both human and object), independent visual cues from the object and the human RoIs, and the fine-grained body part level features. Likewise, Xu *et. al.* [45] use the human pose features in conjunction with the gaze estimates to discover human intentions, which are then used for HOI detection. Since the HOI problem is well-suited for graph-based representations, Graph Convolutional Networks have been regularly used to model the interactions. In this line of work, Xu *et. al.* [46] propose to deal with long-tail HOI categories by modeling underlying regularities among verbs and objects. They do so by constructing a knowledge graph and enforcing similarity of graph embeddings derived from a GCN with visual feature embeddings derived from a CNN using a triplet-loss. Qi *et. al.* [32] propose GPNN, a method that uses an iterative message passing framework on a parse graph comprising of verbs and objects as nodes. Our work is inspired by graph based methods in that we represent humans and objects as graph nodes and learn their interactions based on the image-based node features.

HOI from Video: The HOI labels predicted in this task are typically indicative of an activity spanning over a period of time. Therefore, utilizing temporal cues in a video setting is naturally expected to provide important insights on the interactions and thereby benefit the HOI detection. Albeit less, there have also been significant contributions towards research on HOI detection in videos, mostly on the CAD-120 dataset. Koppula *et. al.* [15] proposed the dataset and introduced an MRF base formulation for handling spatio-temporal requirements. The authors hand-crafted a set of features for humans (pose, displacement of joints, *etc.*) and objects (3D centroids, transforms of SIFT matches between adjacent frames, *etc*). Instead of being used at the frame-level, these features, put together, represented a video segment as a whole. Since then, most existing methods (deep learning and traditional methods alike) work on the same segment level features. Qi *et. al.* [32] extend their GPNN method for videos and construct a parse graph for every video segment using the segment level features to initialize the node and edge features in their parse graph. Likewise, Jain *et. al.* [12] design a spatio-temporal graph for performing structured predictions on a video consisting of multiple segments. Kopulla *et. al.* [16] present ATCRF - a CRF based approach that models anticipatory trajectories of objects and humans.

While there have been remarkable improvements over the years, we submit that there are two major areas for improvement. Firstly, avoiding over-dependence on such hand-crafted features, since the above approaches limit the scope for in-the-wild HOI detections. Such over-dependence has been averted in both textual [2] and image [18] domains and we take inspiration from such works. More often than not, the 3D poses or 3D centroids of objects (used as features) are either not available or are too erroneously estimated to be simply plugged into a model trained on hand-crafted features. Secondly, all the existing methods model temporal relations only between multiple *segments* of a video. This may be, partly, because the hand-crafted features discussed above are defined for a segment as a whole. We believe that there is scope for exploring temporal cues even at a more fine-grained level, *viz.*, frame-level. Using image-based features facilitates the same.

We, therefore, propose an approach to model HOI relevant spatial-structures from every frame of a segment and further design a temporal aggregation regime using these frame level structures. Again, such aggregation strategies have provided to be extremely effective in other domains such a entity-linking [17, 19]. Deep-learning based computer vision models have enough representation power to be able to extract meaningful visual features from images or videos. Thus, our primary intent is to construct a model which can effectively learn hierarchical HOI embeddings at a fine-grained frame level as well as at a coarser segment level, using only visual attributes, and set a new baseline for human-object interaction detection in videos.

3 OUR APPROACH: LIGHTEN

In this section, we present our method, LIGHTEN (Learning Interactions using Graphs and Hierarchical TEmporal Networks) for HOI detection on video. The HOI information in the videos can be dealt with at two levels of granularity. The first, and the coarser, granularity corresponds to viewing the video as a sequence of segments, with each segment representing an atomic interaction. For example, a video may include a sequence of segments such as: *reaching* for a jar, *opening* the jar, and *placing* the jar back. The second, and finer, granularity corresponds to dissecting each segment into its constituent frames. Lastly, the visual features at frame level provide crucial spatial cues about the possible interactions.

In LIGHTEN, we attempt to exploit these well defined constructs and put them under consideration when choosing the architecture. The overall pipeline of LIGHTEN is illustrated in Figure 2.

3.1 The Proposed Learning Architecture

Given an input video $\mathcal{I} = \{I_1, I_2, \ldots, I_T\}$ consisting of T frames such that the video includes a single human and N objects, our task is to regress human subactivities (placing, opening, *etc.*), $H = \{H_0, H_1, \ldots, H_M\}$ for the human and object affordances (placable, openable *etc.*), $O = \{O_{0,0}, O_{0,1}, \ldots, O_{N,M}\}$ for each of the N objects and M segments in the video. To this end, we propose a pipeline consisting of three stages: (i) the spatial subnet, (ii) the frame-level temporal subnet, and (iii) the segment-level temporal subnet.

The spatial subnet feeds on an input frame I_t and learns a set of embeddings $\phi_t \in \mathbb{R}^{D_{emb}}$ for each human and $\theta_{n,t} \in \mathbb{R}^{D_{emb}}$ for each object. These per-frame, spatial embeddings are then fed to the *frame-level* temporal subnet that churns out the corresponding spatio-temporal embeddings, $\Phi_t \in \mathbb{R}^{D_{emb}}$ and $\Theta_{n,t} \in \mathbb{R}^{D_{emb}}$, while also providing initial estimates of H_m and $O_{n,m}$, where m corresponds to the segment index, and n corresponds to the object index. The frame-level spatio-temporal embeddings are then consolidated for each segment using an attention mechanism to produce A^{Φ}_{m} and $A^{\Theta}_{n,m}$, and passed on to *segment-level* temporal subnet that produces the final outputs for the subactivity and affordance estimates.

To the best of our knowledge, LIGHTEN is the first approach to detection of human-object interactions from videos that is completely pivoted on end-to-end learning. On the contrary, majority of prior work [12, 16, 32] has dealt with the problem only at the segment level. Furthermore, previous work has derived spatial features not from the raw images, but from the ground-truth data like

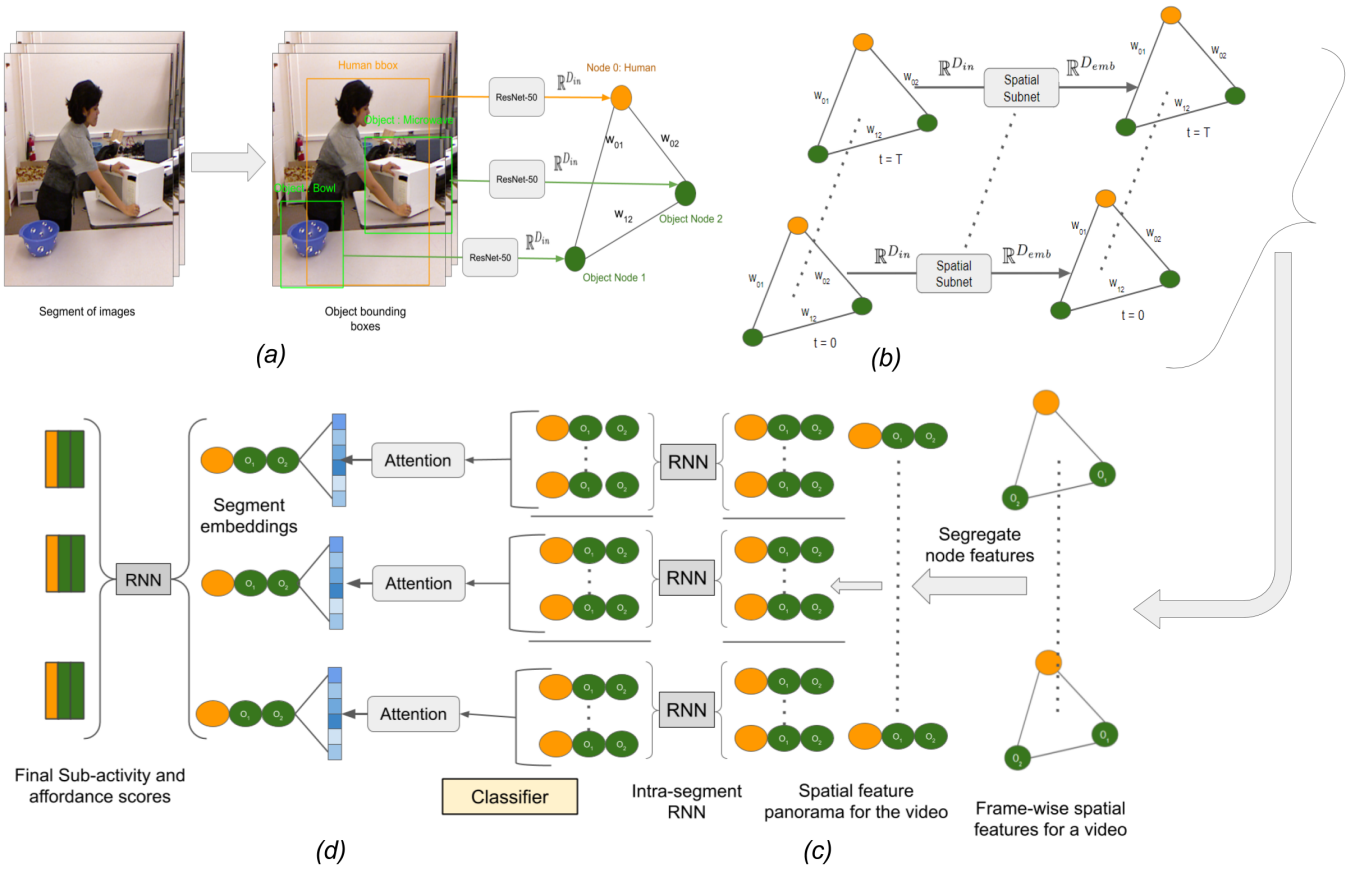

Figure 2: Overall pipeline in LIGHTEN. Given an input video segment with T frames and bounding box coordinates of the humans and objects in every frame, we (a) first extract the visual features from ResNet-50. (b) These features are then processed in a per-frame fashion by a Spatial Subnet. (c) The graph structure is disentangled and temporal cues between frames in a segment are learnt from spatial features. (d) The frame-wise features are summarised into segment embeddings using attention mechanism and refined using inter-segment relations, to regress the human subactivities and object affordances.

depth of the objects, pose of the human and objects, *etc.* It is easy to see that such a construction prohibits its use on any video for which depth information is unavailable. Next, we now elaborate on each step of the pipeline.

3.2 Spatial Subnet

As just discussed, the sole job of the spatial subnet is to learn features relevant to the spatial ordering of the objects and the human. We model this task in a Graph Convolutional Network (GCN) setting which lends itself naturally to the task at hand. We define the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the nodes $\mathcal{V} = \{1, 2, \ldots, N+1\}$ correspond to N objects and one human and $\mathcal{E} = (p, q) \in \mathcal{V} \times \mathcal{V}$.

We extract the node features $x_{v,t} \in \mathbb{R}^{D_{in}}$ corresponding to the v^{th} node (human/object) of the t^{th} frame by feeding the corresponding image crop $I_{v,t}$ to an off-the-shelf feature extractor F. Formally, $x_{v,t} = F(I_{v,t})$. The edge weights are initialized to be 1 for human-object edges and 0 for the rest. The adjacency matrix is dynamically learnt while training the Spatial Subnet.

A major challenge in GCN based formulation is to account for variability in the number of nodes across segments in a video. For example, a video may include the following segments: picking a bowl (1 object), moving the bowl (1 object), putting the bowl in the microwave (2 objects). Typically, this number varies from two nodes to six nodes.

A trivial solution would be to design the GCN with a maximum number of nodes (six, in this case), initialize the unused nodes with zeros, and expect the network to learn to recognize the dummy nodes. This, however, leads to inferior results. To alleviate this issue, the network is designed to inherently learn course-corrections to the adjacency matrix. As depicted in Figure 3, every graph-convolution layer is followed by an update of the adjacency matrix which involves addition of the following two refinement components to the base adjacency matrix A. The first component is a learnable additive matrix, B that is learnt during the training process. The second component is a data-driven additive matrix, C that is estimated uniquely for every input. This formulation has

been inspired by the Adaptive Graph Convolution Network proposed in [38]. However, unlike [38], we do not operate in the time dimension at the level of the GCN.

Formally, the Spatial Subnet, S transforms the features corresponding to the t^{th} frame as $\phi_t = S(x_{v,t})$ if v is a human node and $\theta_t = S(x_{v,t})$ if v corresponds to an object node. At the end of the Spatial Subnet, the network produces an intermediate feature set in $\mathbb{R}^{T\times(N+1)\times D_{emb}}$ space.

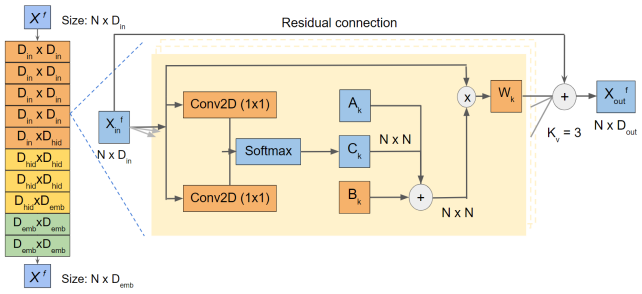

Figure 3: Architecture of Spatial Subnet. Each block augments the adjacency matrix by a learnable correction, B, and a data-dependent course-correction, C. A residual connection is added to facilitate faster training of the model

3.3 Frame-level Temporal Subnet

Once the per-frame spatial features for the graph are extracted, (in the case of video data such as CAD-120) we process the graph features in time dimension, thus providing a feature-panorama of the entire segment. As discussed earlier, temporal reasoning occurs in two granularities - frame level and segment level. It is at this stage that we dis-integrate the graph structure of the network and construct individual feature sets for each node, aggregated over time. These frame-level embeddings are subjected to a bidirectional Recurrent Neural Network (RNN) which produces two outputs for every frame:

For human nodes, given the input embeddings $\phi_t \in \mathbb{R}^{T\times N\times D_{emb}}$, the frame-level bidirectional-RNN outputs the estimates of human subactivity, $H_{m,t}$, and updates the recurrent embedding, $\Phi_t \in \mathbb{R}^{D_{emb}}$ for frame t in segment m. Note, that while the learnt embeddings are further fed into the segment-level subnet, we also use them to classify subactivities and affordances for each frame to facilitate stronger supervision. For object nodes, we concatenate human node features along with the object node features and feed it to the frame-level RNN which outputs the estimates of object affordances $O_{n,m,t}$ and updates the corresponding recurrent embeddings, $\Theta_{n,t} \in \mathbb{R}^{D_{emb}}$

The aggregated activity and affordance classification scores at frame level are computed by taking a summation of the sequential frame-wise scores output by the RNN. Formally, the frame-level subactivity prediction can be written as: $H_m = softmax(\sum_t H_{m,t})$

One key driver behind this form of segregated temporal aggregation for each object, as opposed to joint inference across all objects and humans is the variability in the number of objects in the scene. As such, we leave the job of inter-object relationship discovery to the spatial subnet and only exploit human-object correlations while making the temporal predictions.

Table 1: A comparison of our approach with the existing methods. Note that unlike ours, all the methods that we compare with have been trained on hand-crafted features derived from the ground-truth spatial attributes including 3D human pose, object centroids. We obtain the state-of-the-art results in both subactivity, affordance detection tasks while learning the embeddings from RGB data. Seg-RNN corresponds to segment-level RNN

Method	F1 Score in %	
	Sub-activity	Object Affordance
ATCRF [16]	80.4	81.5
S-RNN [12]	83.2	88.7
S-RNN (multi-task) [12]	82.4	91.1
GPNN [32]	88.9	88.8
LIGHTEN w/o Seg-RNN	**85.9**	**88.9**
LIGHTEN (full model)	**88.9**	**92.6**

Loss Functions: We subject both the classifiers to standard Cross-Entropy losses $\mathcal{L}_h$ and $\mathcal{L}_o$, The overall loss is a weighted sum of the two losses and can be written as:

$$\mathcal{L} = \mathcal{L}_h + \lambda\mathcal{L}_o$$

3.4 Segment-level Temporal Subnet

The previous subnet learns intra-segment temporal relations, but does not utilize the temporal information from the previous segments of the video, thus lacking wider context. With the segment-level subnet, we aim to learn inter-segment temporal cues by leveraging the context from previous segments of the video. We use another RNN to model these relations.

Attention Mechanism: The input to the segment-level RNN is a sequence of embeddings, A^{Φ}_m, corresponding to each segment for human nodes. We extract A^{Φ}_m by subjecting the frame-level embeddings, $\Phi_{m,t}$ to an attention network that produces a single embedding for a segment. Formally, $A^{\Phi}_m = \sum_t a_t * \Phi_{m,t}$, where a_t are the attention weights produced by a Multi-Layered Perceptron (MLP). Similar construction follows for the derivation of A^{Θ}_m.

An alternative to this approach could have been to use the embedding corresponding to the last time step, $\Phi_{m,T}$ as the input to the segment-level RNN. While it works well, we observed superior performance with attention-guided mechanism.

The summarized sequence of segment embeddings is finally processed by an RNN, to leverage temporal dependencies from the previous segments for predicting human subactivity and object affordances for the current segment.

We use the same loss functions for classifiers at both frame-level and segment-level.

3.5 Implementation Details

We now discuss implementation details from two vantage points: model and training.

Model: Since the number of frames in a video segment may vary significantly, we uniformly sample a fixed number of frames, T, from the segment (for our experiments on CAD-120 dataset, we use T=20). We extract the RoI crops from each frame and reshape them to a fixed size of 224×224×3 (input dimension for ResNet). For our experiments, we explore the usage of ResNet-34, ResNet-50 [10] as the feature extractors that produces 512-dimensional (2048 for ResNet-50) features for every node of the graph. Since we have limited data, we use the pre-trained ResNet features. In order to incorporate the information on positioning of humans and objects, we append normalized bounding box coordinates of human/objects to their respective visual node features. We use a hidden, output feature dimensions of 512 for the graph convolutional network of Spatial Subnet.

Training: We use the PyTorch deep learning framework for implementing LIGHTEN. During training, we set $\lambda = 2$ for the overall loss. We use the Adam [14] optimizer with initial learning rate of 2×10^{-5}, learning rate decay factor of 0.8, and decay step size of 10 epochs. We train LIGHTEN for a total of 300 epochs on Nvidia RTX 2080Ti GPU. We performed a hyper-parameter sweep to empirically obtain these configurations. The entire model is trained in two steps. Firstly, the model up to frame-level temporal subnet is trained by aggregating classification scores from the T frames of the segment. Finally, the entire model including the segment-level subnet, is trained in an end-to-end fashion, after initializing the parameters from the pre-trained frame-level model.

4 EXPERIMENTS

4.1 Datasets

We evaluate LIGHTEN for the task of Human-Object Interaction detection on two datasets, *viz*., i) CAD-120 [15] ,and ii) V-COCO [9].

CAD-120: The CAD-120 dataset is a video dataset with 120 RGB-D videos of 4 subjects performing 10 daily indoor activities (*e.g., making cereal, microwaving food*). Each activity is a sequence of video segments involving finer-level activities. In each video segment, the human is annotated with an activity label from a set of 10 sub-activity classes *(e.g., reaching, pouring)* and each object is annotated with an affordance label from a set of 12 affordance classes (*e.g., pourable, movable*). The frame-length of each segment ranges from 22 to a little over 150 frames.

The metrics used for evaluating LIGHTEN on the human-object interaction tasks of CAD-120 dataset are: i) sub-activity F1-score, and ii) object affordance F1-score computed for human sub-activity classification and object affordance classification. The dataset, in addition to providing the images and HOI annotations, additionally provides depth maps, 3D pose information and segment-level hand-crafted spatial features. We do not make use of any additional data except the 2D bounding box of the objects and humans, and aim to learn the segment embeddings from RGB data only.

V-COCO: Crafted as a subset of the MS-COCO [22] dataset, V-COCO is an image dataset that provides annotations of Action labels for edges between human and object. There are 26 action classes.

4.2 Quantitative Evaluation

4.2.1 Evaluation on the CAD-120 dataset: The performance of LIGHTEN is evaluated in two experimental setups. i) In the first setup, we pick the labels predicted directly from the output sequence at the frame-level subnet. In the second setup, ii) the subactivity and affordances are predicted after incorporating the segment-level RNN. In each of these two experiments, we train LIGHTEN separately for the tasks of HOI detection and HOI anticipation. In all the experiments, the video data we provide as input to LIGHTEN is: i) RGB frames of the video ii) bounding boxes of human and object in the frames of video.

We tabulate the results of our approach in Table 1. As the numbers suggest, we achieve state-of-the-art performance with sub-activity detection F1 score of **88.9** and affordance detection F1 score of **92.6**. we also achieve an F1 score of **76.4** in human sub-activity anticipation task, outperforming previous methods, and an F1 score of **78.8** in affordance anticipation task. To the best of our knowledge, all previous works on the task of human-object interaction in CAD-120, use the hand-crafted features provided by CAD-120 dataset. So we believe that this experiment is the only one which bypasses the usage of the handcrafted features and relies only on 2D video data, while achieving improved performance.
We compare our method against the existing works on CAD-120: ATCRF [16], S-RNN [12], and GPNN [32].

Confusion Matrix: The confusion matrices for both detection and anticipation tasks are displayed at Figure 4. Every row of a confusion matrix indicates the prediction distribution of various node samples of that ground truth class. From the confusion matrix for affordance detection, it is evident that most of the false predictions of object nodes are due to misinterpretation of object as stationary. This is especially prevalent in the affordance class *reachable*, because the human is usually far from the object during the sub-activity *reaching*.

Table 2: A comparison of LIGHTEN on image-based HOI detection on V-COCO dataset

Method	Role mAP score
Gupta et al. [9]	31.8
InteractNet [7]	40.0
GPNN [32]	44.0
Li et al. [21]	48.6
PMFNet [42]	52.0
LIGHTEN for image HOI	38.28

4.2.2 Evaluation on V-COCO dataset. Although our method is designed to leverage temporal cues within a video setting, we test our method on V-COCO dataset by setting T = 1. We observe the role mAP score of 38.28 which, although not close to the state-of-the-art method, achieves well reasonable performance without bells and whistles. We believe that an explanation for the sub-parity of our

results is that in the absence of temporal cues, the spatial GCN is significantly shallower than other works and leads to inferior results. We provide a detailed comparison with other methods in Table 2.

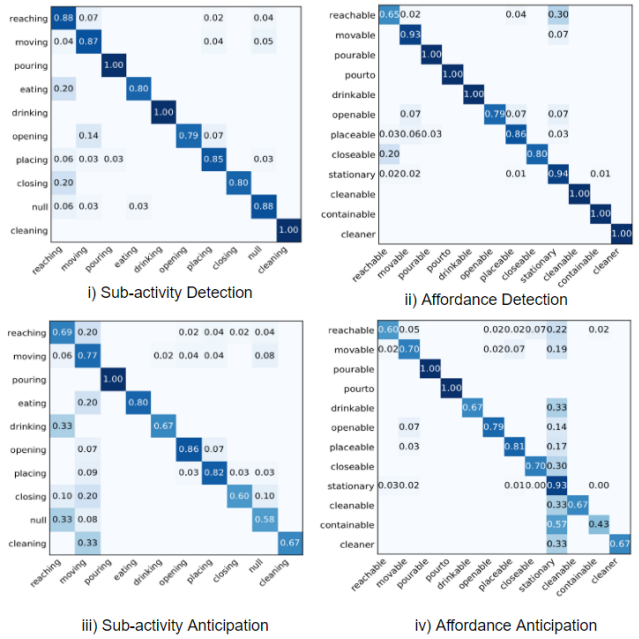

Figure 4: Confusion matrices for human-object interaction detection setting – (i), (ii) – and anticipation setting – (iii), (iv) – on CAD120 dataset. It is worth noting that most of the confusion occurs in visually similar categories, e.g. closing vs. reaching and opening vs. moving

4.3 Qualitative Evaluation

We provide some qualitative evaluation of LIGHTEN on CAD-120 dataset in Figure 5. We see that while the HOI detections have been achieved accurately, there remains ambiguity among some classes during the anticipation task.

Figure 6 demonstrates some positive and negative cases of detection of edge action labels of human-object pairs for test images on V-COCO. In the absence of temporal context, the method resorts to associating visual cues to spatial cues, thus not being able to disambiguate whether a person is *sitting* on a car or *looking* at the same car.

4.4 Ablation Study

We now discuss the contributions of various components to the final performance and their relevance to Human-Object Interaction detection.

4.4.1 Role of Graph Convolutions in Spatial Subnet: Firstly, to verify the effectiveness of spatial graph convolution module, we designed an experiment where the image features from the backbone are directly passed to the frame-level model. We observed a significant degradation in performance in the absence of spatial

Table 3: A comparison of LIGHTEN on anticipation task. Our approach achieves state-of-the-art results on human subactivity anticipation whereas performs competitively on object affordance anticipation.

Method	F1 Score in %	
	Sub-activity	Object Affordance
ATCRF [16]	37.9	36.7
S-RNN [12]	62.3	80.7
S-RNN (multi-task) [12]	65.6	80.9
GPNN [32]	75.6	81.9
LIGHTEN w/o Segment-level subnet	**73.2**	**77.6**
LIGHTEN (full model)	**76.4**	**78.8**

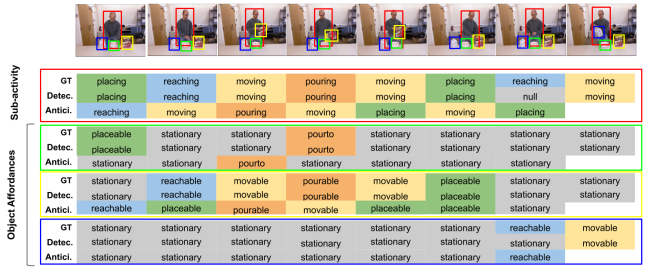

Figure 5: Human Object Interaction Detection and Anticipation results on a video of activity "Making cereal" from the CAD-120 dataset.The nodes here are the human and three objects: i) bowl ii) milk iii) box. The object affordance predictions in the figure are for the objects in this order from top to bottom. Predictions are highlighted a border of same color (red for human, green for bowl, blue for milk, and yellow for box) as the human/object's bounding box in image. We show predictions for 8 segments of the video. The anticipation labels shown along with each segment are the labels anticipated for the upcoming segment.

GCN. While exploring variants of Graph Convolutional Networks, we also explored using a vanilla GCN network with basic graph convolution (GCN) layers as a baseline. As an extension to the basic GCN, we add a residual connection, similar to [10], which allows the input features to retain their initial behaviour. Using a residual connection brings an improvement in performance of GCN, as illustrated in Table 4. Further adding adaptive and data-dependent components to adjacency matrix, in a fashion similar to [38], also improves subactivity and affordance prediction, largely due to the ability to learn the inter-node edge weights.

4.4.2 Role of human node features in affordance prediction: In the temporal subnet, we concatenate human node features along with object node features for the frame and segment level RNNs. We observed significant improvement in performance on object affordance detection (88.6% vs 84.6%) due to human node features. This improvement can be attributed to the high correlation between the human sub-activity and affordances of active objects (objects which are not stationary).

Table 4: Ablation experiments of the impact of design choices on subactivity and object affordance detection. Seg-RNN refers to segment-level RNN and vanilla GCN refers to GCN without adjacency matrix refinement.

Experiment	Human Subactivity	Object Affordance
LIGHTEN w/o seg-RNN w/o spatial GCN	61.5	78.6
LIGHTEN w/o seg-RNN with vanilla GCN block w/o residual connections	70.3	61.3
LIGHTEN w/o seg-RNN with vanilla GCN block with residual connections	79.3	83.1
LIGHTEN w/o seg-RNN with MLP for frame-level temporal learning	84.1	85.0
LIGHTEN w/o seg-RNN w/o appending human node features to object nodes	85.2	84.6
LIGHTEN w/o seg-RNN	85.9	88.9
LIGHTEN w/o attention	83.5	86.1
LIGHTEN w/o seg-RNN with MLP for segment level temporal function	89.7	90.5
Seg-RNN on hand-crafted features	85.3	91.6
LIGHTEN	88.9	92.6

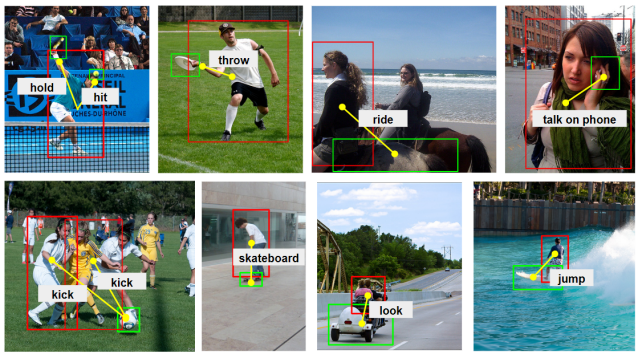

Figure 6: Detections of human-object action labels in test images of VCOCO. We report our failure cases on the last two images (bottom right). The rest are correct predictions.

4.4.3 Role of RNN in frame-level temporal subnet: As a baseline for classification at frame-level subnet, we experimented with alternative temporal aggregation models. Specifically, we built an MLP network to obtain classification scores from spatial features concatenated across temporal dimension for each node separately. However, due to higher parameter count in MLP network, the model is prone to over-fitting, and therefore has a lower performance, which is evident from Table 4.

4.4.4 Role of segment-level temporal learning: Even though subactivity and affordance labels are predicted for every single segment, there are significant inter-dependencies between the activity in a segment and activities in previous segments. As an illustrative example, in the following sequence of three segments in a video: *reaching* for a jar, *moving* the jar, and *placing* the jar back, knowledge on the activities in first two segments can greatly improve the prediction of activity in the third segment. Using a temporal sequence processing network like an RNN after the frame-level aggregation step leverages these inter-segment dependencies and achieves a significant improvement in performance as compared to prediction at frame-level temporal subnet.

4.4.5 Role of attention-mechanism in computing segment embedding: We implemented two simpler baseline approaches to evaluate the use of attention weighting for frames. These approaches include i) using features corresponding to last frame in the output sequence of RNN ii) stitching the features across frames and regressing a segment embedding using MLP. Using the embedding corresponding to the last frame limits the representation power of the segment-level embedding Φ_m. Using an MLP has the disadvantage of over-fitting and has an impact on object affordance detection as evident from the Table 4.

4.4.6 Evaluating the feature learning process: To measure the effectiveness of the hierarchical learning mechanism, we design an experiment where we feed the hand-crafted, segment-level features to segment-level RNN, instead of the visual embeddings learnt by the attention mechanism. The learnt visual features achieve a better performance than the hand crafted features, particularly for the more difficult case of human subactivity detection (85.3% vs 88.9%), thereby justifying the effectiveness of the proposed method in capturing the spatio-temporal relations from RGB video data.

5 CONCLUSION

In this paper, we proposed a two-step hierarchical approach for identifying Human-Object Interaction in videos. In the first step, we model the local interactions between humans and objects at a frame-level, while in the second step, we generate a segment-level embedding using the frame-level embeddings, and then refine them using the embeddings from previous segments. The embeddings are modelled through a graph structure, where the subject and object serve as nodes in a scene. Our approach is easily extendable to other videos for the task of HOI, where depth information and 3D pose information is not available. Our approach sets a new benchmark for Human-Object Interaction detection in videos with visual information.

ACKNOWLEDGEMENTS

We are grateful to IBM Research, India (specifically the IBM AI Horizon Networks - IIT Bombay initiative) for their support and sponsorship. Rishabh Dabral has also been supported by Qualcomm Innovation Fellowship 2019.

REFERENCES

[1] Y.W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. 2015. HICO: A benchmark for recognizing human-object interactions in images. In *ICCV*.

[2] Oishik Chatterjee, Ganesh Ramakrishnan, and Sunita Sarawagi. 2020. Robust Data Programming with Precision-guided Labeling Functions. In *AAAI*.

[3] R. Dabral, N. B. Gundavarapu, R. Mitra, A. Sharma, G. Ramakrishnan, and A. Jain. 2019. Multi-Person 3D Human Pose Estimation from Monocular Images. In *3DV*.

[4] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. 2018. Learning 3D Human Pose from Structure and Motion. In *ECCV*.

[5] Navneet Dalal and Bill Triggs. 2005. Histograms of Oriented Gradients for Human Detection. In *CVPR*.

[6] V. Delaitre, J. Sivic, and I. Laptev. 2011. Learning person-object interactions for action recognition in still images. In *NIPS*.

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

[8] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. DensePose: Dense Human Pose Estimation In The Wild. In *CVPR*.

[9] Saurabh Gupta and Jitendra Malik. 2015. Visual Semantic Role Labeling. In *arXiv preprint arXiv:1505.04474*.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

[11] J.F. Hu, W.S. Zheng, J. Lai, S. Gong, and T. Xiang. 2013. Recognising human-object interaction via exemplar based modelling. In *ICCV*.

[12] A. Jain, A.R. Zamir, S. Savarese, and A Saxena. 2016. Structural-RNN: Deep learning on spatio-temporal graphs. In *CVPR*.

[13] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2018. End-to-end Recovery of Human Shape and Pose. In *CVPR*.

[14] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimizations. In *ICLR*.

[15] H.S. Koppula, R. Gupta, and A. Saxena. 2013. Learning human activities and object affordances from RGB-D videos. In *The International Journal of Robotics Research*.

[16] H.S. Koppula and A. Saxena. 2016. Anticipating human activities using object affordances for reactive robotic response. In *TPAMI*.

[17] Ashish Kulkarni, Kanika Agarwal, Pararth Shah, Sunny Raj Rathod, and Ganesh Ramakrishnan. 2016. Learning to Collectively Link Entities. In *Proceedings of the 3rd IKDD Conference on Data Science, CODS*.

[18] Ashish Kulkarni, Narasimha Raju Uppalapati, Pankaj Singh, and Ganesh Ramakrishnan. 2018. An Interactive Multi-Label Consensus Labeling Model for Multiple Labeler Judgments. In *AAAI*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press.

[19] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. ACM SIGKDD, John F. Elder IV, Françoise Fogelman-Soulié, Peter A. Flach, and Mohammed Javeed Zaki (Eds.).

[20] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 2019. 360-Degree Textures of People in Clothing from a Single Image. In *3DV*.

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

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *ECCV*.

[23] Chenchen Liu, Yang Jin, Kehan Xu, Guoqiang Gong, and Yadong Mu. 2020. Beyond Short-Term Snippet: Video Relation Detection with Spatio-Temporal Global Context. In *CVPR*.

[24] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. 2020. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In *CVPR*.

[25] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. 2020. XNect: Real-time Multi-Person 3D Motion Capture with a Single RGB Camera. *ACM Transactions on Graphics*.

[26] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege's Principle. In *CVPR*.

[27] Arsha Nagrani, Chen Sun, David Ross, Rahul Sukthankar, Cordelia Schmid, and Andrew Zisserman. 2020. Speech2Action: Cross-Modal Supervision for Action Recognition. In *CVPR*.

[28] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked hourglass networks for human pose estimation. In *ECCV*.

[29] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. 2018. Neural Body Fitting: Unifying Deep Learning and Model Based Human Pose and Shape Estimation. In *3DV*.

[30] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. 2020. TailorNet: Predicting Clothing in 3D as a Function of Human Pose, Shape and Garment Style. In *CVPR*.

[31] O. Pele and M. Werman. 2008. A linear time histogram metric for improved sift matching. In *ECCV*.

[32] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. 2018. Learning Human-Object Interactions by Graph Parsing Neural Networks. In *ECCV*.

[33] Xufeng Qian, Yueting Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, and Jun Xiao. 2019. Video Relation Detection with Spatio-Temporal Graph. In *ACM MM*.

[34] Tanmay Randhavane, Aniket Bera, Kyra Kapsaskis, Rahul Sheth, Kurt Gray, and Dinesh Manocha. 2019. EVA: Generating Emotional Behavior of Virtual Agents Using Expressive Features of Gait and Gaze. In *ACM Symposium on Applied Perception*.

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*.

[36] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating Objects and Relations in User-Generated Videos. In *ICMR*.

[37] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. 2017. Video Visual Relation Detection. In *ACM MM*.

[38] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In *CVPR*.

[39] Xu Sun, Tongwei Ren, Yuan Zi, and Gangshan Wu. 2019. Video Visual Relation Detection via Multi-modal Feature Fusion. In *ACM MM*.

[40] Yao-Hung Hubert Tsai, Santosh Divvala, Louis-Philippe Morency, Ruslan Salakhutdinov, and Ali Farhadi. 2019. Video Relationship Reasoning using Gated Spatio-Temporal Energy Graph. In *CVPR*.

[41] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. 2017. Learning From Synthetic Humans. In *CVPR*.

[42] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. 2019. Pose-aware Multi-level Feature Network for Human Object Interaction Detection. In *ICCV*.

[43] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. 2018. Repulsion Loss: Detecting Pedestrians in a Crowd. In *CVPR*.

[44] Yunyang Xiong, Hyunwoo J. Kim, and Vikas Singh. 2019. Mixed Effects Neural Networks (MeNets) With Applications to Gaze Estimation. In *CVPR*.

[45] Bingjie Xu, Junnan Li, Yongkang Wong, Mohan S Kankanhalli, and Qi Zhao. 2018. Interact as you intend: Intention driven human-object interaction detection. In *arXiv preprint arXiv:1808.09796*.

[46] B. Xu, Y. Wong, J. Li, Q. Zhao, and M. S. Kankanhalli. 2019. Learning to Detect Human-Object Interactions With Knowledge. In *CVPR*.

[47] B. Yao and L. Fei-Fei. 2010. Grouplet: A structured image representation for recognizing human and object interactions.. In *CVPR*.

[48] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. 2011. Human Action Recognition by Learning Bases of Action Attributes and Parts. In *ICCV*.

[49] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. 2020. Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition. In *CVPR*.

[50] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. 2018. Occlusion-aware R-CNN: Detecting Pedestrians in a Crowd. In *ECCV*.