# Action Genome: Actions as Compositions of Spatio-temporal Scene Graphs

Jingwei Ji  Ranjay Krishna  Li Fei-Fei  Juan Carlos Niebles
Stanford University
{jingweij, ranjaykrishna, feifeili, jniebles}@cs.stanford.edu

## Abstract

*Action recognition has typically treated actions and activities as monolithic events that occur in videos. However, there is evidence from Cognitive Science and Neuroscience that people actively encode activities into consistent hierarchical part structures. However, in Computer Vision, few explorations on representations that encode event partonomies have been made. Inspired by evidence that the prototypical unit of an event is an action-object interaction, we introduce Action Genome, a representation that decomposes actions into spatio-temporal scene graphs. Action Genome captures changes between objects and their pairwise relationships while an action occurs. It contains* $10K$ *videos with* $0.4M$ *objects and* $1.7M$ *visual relationships annotated. With Action Genome, we extend an existing action recognition model by incorporating scene graphs as spatio-temporal feature banks to achieve better performance on the Charades dataset. Next, by decomposing and learning the temporal changes in visual relationships that result in an action, we demonstrate the utility of a hierarchical event decomposition by enabling few-shot action recognition, achieving* $42.7\%$ *mAP using as few as* $10$ *examples. Finally, we benchmark existing scene graph models on the new task of spatio-temporal scene graph prediction.*

## 1. Introduction

Video understanding tasks, such as action recognition, have, for the most part, treated actions and activities as monolithic events [8, 38, 66, 87]. Most recent models proposed have resorted to end-to-end predictions that produce a single label for a long sequence of a video [10, 23, 31, 69, 72] and do not explicitly decompose events into a series of interactions between objects. On the other hand, image-based structured representations like scene graphs have cascaded improvements across multiple image tasks, including image captioning [2], image retrieval [36, 64], visual question answering [35], relationship modeling [41] and image generation [34]. The scene graph representation, introduced in Visual Genome [43], provides a scaffold that allows vision models to tackle complex inference tasks by breaking scenes into its corresponding objects and their visual relationships. However, decompositions for temporal events have not been explored much [50], even though representing events with structured representations could lead to more accurate and grounded action understanding.

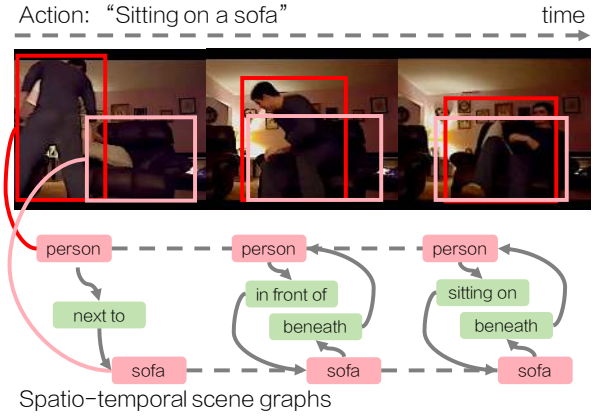


Figure 1: We present Action Genome: a representation that decomposes actions into spatio-temporal scene graphs. Inspired by hierarchical bias theory [84] and event segmentation theory [44], Action Genome provides the scaffold to study the dynamics of actions as relationships between people and objects. This decomposition also allow us to improve action recognition, enable few-shot action detection, and introduce spatio-temporal scene graph prediction.

Meanwhile, in Cognitive Science and Neuroscience, it has been postulated that people segment events into consistent groups [5, 6, 55]. Furthermore, people actively encode those ongoing activities in a hierarchical part structure — a phenomenon referred to as hierarchical bias hypothesis [84] or event segmentation theory [44]. Let's consider the action of "sitting on a sofa". The person initially starts off next to the sofa, moves in front of it, and finally sits atop it. Such decompositions can enable machines to predict future and past scene graphs with objects and relationships as an action occurs: we can predict that the person is about to sit on

Table 1: A comparison of Action Genome with existing video datasets. Built upon Charades [66], Action Genome is the first large-scale video database providing both action labels and spatio-temporal scene graph labels.

| Dataset | Video hours | # videos | # action categories | Objects | | | | Relationships | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | annotated | localized | # categories | # instances | annotated | localized | # categories | # instances |
| ActivityNet [8] | 648 | 28K | 200 | | | - | - | | | - | - |
| HACS Clips [87] | 833 | 0.4K | 200 | | | - | - | | | - | - |
| Kinetics-700 [9] | 1794 | 650K | 700 | | | - | - | | | - | - |
| AVA [26] | 108 | 504K | 80 | | | - | - | ✓ | | 49 | - |
| Charades [66] | 82 | 10K | 157 | ✓ | | 37 | - | | | - | - |
| EPIC-Kitchen [15] | 55 | - | 125 | ✓ | | 331 | - | | | - | - |
| DALY [75] | 31 | 8K | 10 | ✓ | ✓ | 41 | 3.6K | | | - | - |
| CAD120++ [91] | 0.57 | 0.5K | 10 | ✓ | ✓ | 13 | 64K | ✓ | ✓ | 6 | 32K |
| **Action Genome** | 82 | 10K | 157 | ✓ | ✓ | 35 | **0.4M** | ✓ | ✓ | 25 | **1.7M** |

the sofa when we see them move in front of it. Similarly, such decomposition can also enable machines to learn from few examples: we can recognize the same action when we see a different person move towards a different chair. While that was a relatively simple decomposition, other events like "playing football", with its multiple rules and actors, can involve multifaceted decompositions. So while such decompositions can provide the scaffolds to improve vision models, how is it possible to correctly create representative hierarchies for a wide variety of complex actions?

In this paper, we introduce Action Genome, a representation that decomposes actions into spatio-temporal scene graphs. Object detection faced a similar challenge of large variation within any object category. So, just as progress in 2D perception was catalyzed by taxonomies [56], partonomies [57], and ontologies [43, 79], we aim to improve temporal understanding with Action Genome's partonomy. Going back to the example of "person sitting on a sofa", Action Genome breaks down such actions by annotating frames within that action with scene graphs. The graphs captures both the objects, `person` and `sofa`, and how their relationships evolve as the actions progress from ⟨`person-next to-sofa`⟩ to ⟨`person-in front of-sofa`⟩ to finally ⟨`person-sitting on-sofa`⟩. Built upon Charades [66], Action Genome provides 476K object bounding boxes with 1.72M relationships across 234K video frames with 157 action categories.

Most perspectives on action decomposition converge on the prototypical unit of action-object couplets [44, 50, 63, 84]. Action-object couplets refer to transitive actions performed on objects (e.g. "moving a chair" or "throwing a ball") and intransitive self-actions (e.g. "moving towards the sofa"). Action Genome's dynamic scene graph representations capture both such types of events and as such, represent the prototypical unit. With this representation, we enable the study for tasks such as spatio-temporal scene graph prediction — a task where we estimate the decomposition of action dynamics given a video. We can also improve existing tasks like action recognition and few-shot action detection by jointly studying how those actions change visual relationships between objects in scene graphs.

To demonstrate the utility of Action Genome's event decomposition, we introduce a method that extends a state-of-the-art action recognition model [76] by incorporating spatio-temporal scene graphs as feature banks that can be used to both predict the action as well as the objects and relationships involved. First, we demonstrate that predicting scene graphs can benefit the popular task of action recognition by improving the state-of-the-art on the Charades dataset [66] from 42.5% to 44.3% and to 60.3% when using oracle scene graphs. Second, we show that the compositional understanding of actions induces better generalization by showcasing few-shot action recognition experiments, achieving 42.7% mAP using as few as 10 training examples. Third, we introduce the task of spatio-temporal scene graph prediction and benchmark existing scene graph models with new evaluation metrics designed specifically for videos. With a better understanding of the dynamics of human-object interactions via spatio-temporal scene graphs, we aim to inspire a new line of research in more decomposable and generalizable action understanding.

## 2. Related work

We derive inspiration from Cognitive Science, compare our representation with static scene graphs, and survey methods in action recognition and few-shot prediction.

**Cognitive Science.** Early work in Cognitive Science provides evidence for the regularities with which people identify event boundaries [5, 6, 55]. Remarkably, people consistently, both within and between subjects, carve out video streams into events, actions, and activities [11, 28, 83]. Such findings hint that it is possible to predict when actions begin and end, and have inspired hundreds of Computer Vision datasets, models, and algorithms to study tasks like action recognition [19, 37, 71, 80, 81, 82]. Subsequent Cognitive and Neuroscience research, using the same paradigm, has also shown that event categories form partonomies [28, 60, 83]. However, Computer Vision has done little work in explicitly representing the hierarchical structures of actions [50], even though understanding event partonomies can improve tasks like action recognition.

**Action recognition in videos.** Many research projects have tackled the task of action recognition. A major line of work has focused on developing powerful neural architectures to extract useful representations from videos [10, 23, 31, 69, 72]. Pre-trained on large-scale databases for action classification [8, 9], these architectures serve as cornerstones for downstream video tasks and action recognition on other datasets. To assist more complicated action understanding, another growing set of research explores structural information in videos including temporal ordering [51, 88], object localization [4, 25, 32, 53, 74, 76], and implicit interactions between objects [4, 53]. In our work, we contrast against these methods by explicitly using a structured decomposition of actions into objects and relationships.

Table 1 lists some of the most popular datasets used for action recognition. One major trend of video datasets is providing considerably large amount of video clips with single action labels [8, 9, 87]. Although these databases have driven the progress of video feature representation for many downstream tasks, the provided annotations treat actions as monolithic events, and do not study how objects and their relationships change during actions/activities. In the mean time, other databases have provided more varieties of annotations: AVA [26] localizes the actors of actions, Charades [66] contains multiple actions happening at the same time, EPIC-Kitchen [15] localizes the interacted objects in ego-centric kitchen videos, DALY [75] provides object bounding boxes and upper body poses for 10 daily activities. Still, scene graph, as a comprehensive structural abstraction of images, has not yet been studied in any large-scale video database as a potential representation for action recognition. In this work, we present Action Genome, the first large-scale database to jointly boost research in scene graphs and action understanding. Compared to existing datasets, we provide orders of magnitude more object and relationship labels grounded in actions.

**Scene graph prediction.** Scene graphs are a formal representation for image information [36, 43] in a form of a graph, which is widely used in knowledge bases [13, 27, 89]. Each scene graph encodes objects as nodes connected together by pairwise relationships as edges. Scene graphs have led to many state of the art models in image captioning [2], image retrieval [36, 64], visual question answering [35], relationship modeling [41], and image generation [34]. Given its versatile utility, the task of scene graph prediction has resulted in a series of publications [14, 30, 43, 46, 48, 49, 59, 77, 78, 85] that have explored reinforcement learning [49], structured prediction [16, 40, 70], utilizing object attributes [20, 61], sequential prediction [59], few-shot prediction [12, 17], and graph-based [47, 77, 78] approaches. However, all of these approaches have restricted their application to static images and have not modelled visual concepts spatio-temporally.

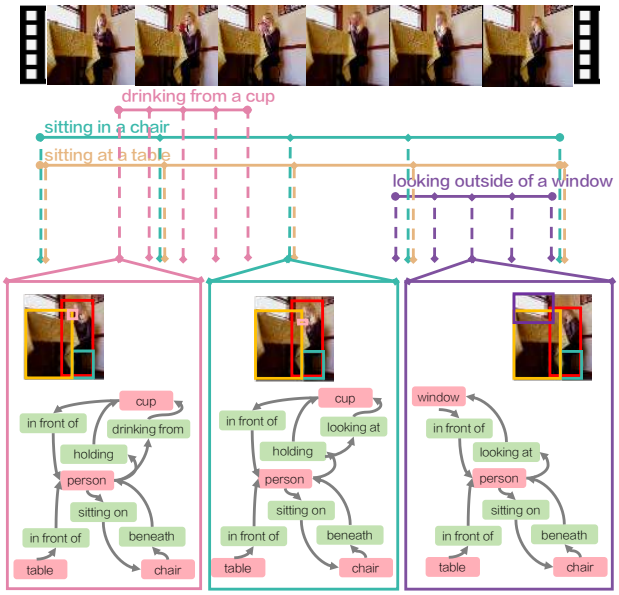


Figure 2: Action Genome's annotation pipeline: For every action, we uniformly sample 5 frames across the action and annotate the person performing the action along with the objects they interact with. We also annotate the pairwise relationships between the person and those objects. Here, we show a video with 4 actions labelled, resulting in 20 ($= 4 \times 5$) frames annotated with scene graphs. The objects are grounded back in the video as bounding boxes.

**Few-shot prediction.** The few-shot literature is broadly divided into two main frameworks. The first strategy learns a classifier for a set of frequent categories and then uses them to learn the few-shot categories [21, 22, 58]. For example, ZSL uses attributes of actions to enable few-shot [58]. The second strategy learns invariances or decompositions that enable few-shot classification [7, 18, 39, 90]. OSS and TARN propose a measurement of similarity or distance measure between video pairs [7, 39], CMN encodes uses a multi-saliency algorithm to encode videos [90], and Proto-GAN creates a prototype vector for each class [18]. Our framework resembles the first strategy because we use the object and visual relationship representations learned using the frequent actions to identify few-shot actions.

## 3. Action Genome

Inspired from Cognitive Science, we decompose events into prototypical action-object units [44, 63, 84]. Each action in Action Genome is representated as changes to objects and their pairwise interactions with the actor/person performing the action. We derive our representation as a temporally changing version of Visual Genome's scene graphs [43]. However, unlike Visual Genome, whose goal was to densely represent a scene with objects and visual re-

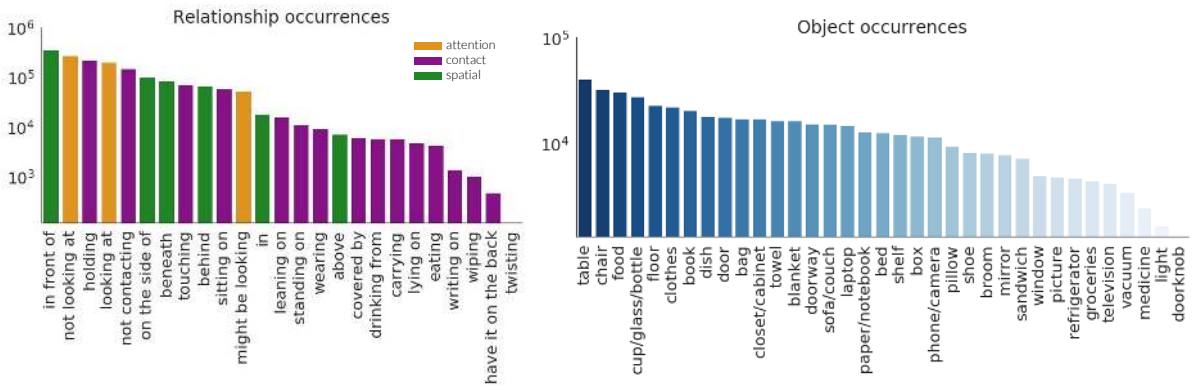


Figure 3: Distribution of (a) relationship and (b) object occurrences. The relationships are color coded to represent attention, spatial , and contact relationships. Most relationships have at least 1k instances and objects have at least 10k instances.

Table 2: There are three types of relationships in Action Genome: attention relationships report which objects people are looking at, spatial relationships indicate how objects are laid out spatially, and contact relationships are semantic relationships involving people manipulating objects.

| attention | spatial | contact | |
|---|---|---|---|
| looking at | in front of | carrying | covered by |
| not looking at | behind | drinking from | eating |
| unsure | on the side of | have it on the back | holding |
| | above | leaning on | lying on |
| | beneath | not contacting | sitting on |
| | in | standing on | touching |
| | | twisting | wearing |
| | | wiping | writing on |

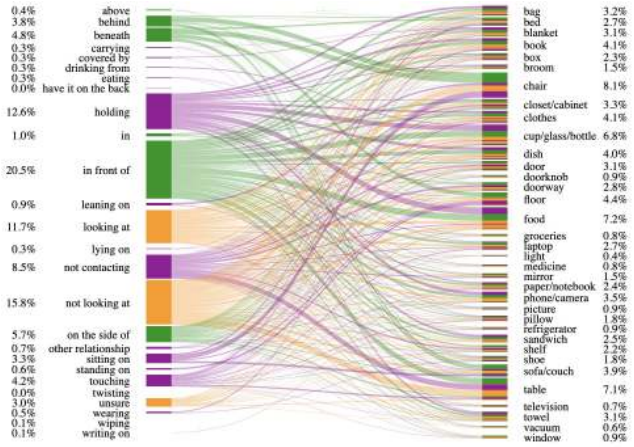


Figure 4: A weighted bipartite mapping between objects and relationships shows that they are densely interconnected in Action Genome. The weights represent percentage of occurrences in which a specific object occurs in a relationship. There are three colors in the graph and they represent the three kinds of relationships: attention (in orange), spatial (in green) and contact (in purple).

lationships, Action Genome's goal is to decompose actions and as such, focuses on annotating only those segments of the video where the action occurs and only those objects that are involved in the action.

**Annotation framework.** Action Genome is built upon the videos and temporal action annotations available in the Charades dataset [66], which contains 157 action classes, 144 of which are human-object activities. In Charades, there are multiple actions that might be occurring at the same time. We do not annotate every single frame in a video; it would be redundant as the changes between objects and relationships occur at longer time scales.

Figure 2 visualizes the pipeline of our annotation. We uniformly sample 5 frames to annotate across the range of each action interval. With this action-oriented sampling strategy, we provide more labels where more actions occur. For instance, in the example, actions "sitting on a chair" and "drinking from a cup" occur together and therefore, result in more annotated frames, 5 from each action. When annotating each sampled frame, the annotators hired were prompted with action labels and clips of the neighboring video frames for context. The annotators first draw bounding boxes around the objects involved in these actions, then choose the relationship labels from the label set. The clips are used to disambiguate between the objects that are actually involved in an action when multiple instances of a given category is present. For example, if multiple "cups" are present, the context disambiguates which "cup" to annotate for the action "drinking from a cup".

Action Genome contains three different kinds of human-object relationships: *attention*, *spatial* and *contact* relationships (see Table 2). Attention relationships indicate if a person is looking at an object or not, and serve as indicators for which object the person is or will interacting with. Spa-

tial relationships describe where objects are relative to one another. Contact relationships describe the different ways the person is contacting an object. A change in contact often indicates the occurrence of an actions: for example, changing from ⟨person - not contacting - book⟩ to ⟨person - holding - book⟩ may show an action of "picking up a book".

It is worth noting that while Charades provides an injective mapping from each action to a verb, it is different from the relationship labels we provide. Charades' verbs are clip-level labels, such as "awaken", while we decompose them into frame-level human-object relationships, such as a sequence of ⟨person - lying on - bed⟩, ⟨person - sitting on - bed⟩ and ⟨person - not contacting - bed⟩.

**Database statistics.** Action Genome provides frame-level scene graph labels for the components of each action. Overall, we provide annotations for 234,253 frames with a total of 476,229 bounding boxes of 35 object classes (excluding "person"), and 1,715,568 instances of 25 relationship classes. Figure 3 visualizes the log-distribution of object and relationship categories in the dataset. Like most concepts in vision, some objects (e.g. table and chair) and relationships (e.g. in front of and not looking at) occur frequently while others (e.g. twisting and doorknob) only occur a handful of times. However, even with such a distribution, almost all objects have at least 10K instances and every relationship as at least 1K instances.

Additionally, Figure 4 visualizes how frequently objects occur in which relationships. We see that most objects are pretty evenly involved in all three types of relationships. Unlike Visual Genome, where dataset bias provides a strong baseline for predicting relationships given the object categories, Action Genome does not suffer the same bias.

## 4. Method

We validate the utility of Action Genome's action decomposition by studying the effect of combining learning spatio-temporal scene graphs with learning to recognize actions. We propose a method, named Scene Graph Feature Banks (SGFB), to incorporate spatio-temporal scene graphs into action recognition. Our method is inspired by recent work in computer vision that uses the information "banks" [1, 45, 76]. Information banks are feature representations that have been used to represent, for example, object categories that occur in the video [45], or even include where the objects are [1]. Our model is most directly related to the recent long-term feature banks [76], which accumulates features of a long video as a fixed size representation for action recognition.

Overall, our SGFB model contains two components: the first component generates spatio-temporal scene graphs while the second component encodes the graphs to predict

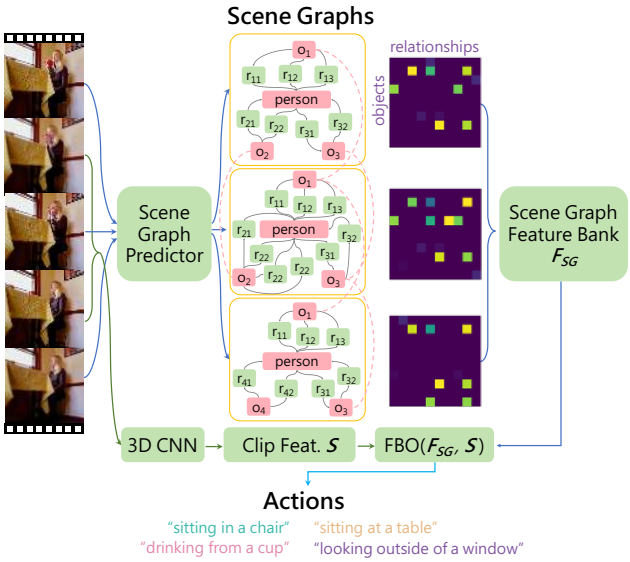


Figure 5: Overview of our proposed model, SGFB, for action recognition using spatio-temporal scene graphs. SGFB predicts scene graphs for every frame in a video. These scene graphs are converted into features representations that are then combined using methods similar to long-term feature banks [76]. The final representation is merged with 3D CNN features and used to predict action labels.

action labels. Given a video sequence $v = \{i_1, i_2, \ldots, i_N\}$, the aim of traditional multi-class action recognition is to assign multiple action labels to this video. Here, $v$ represents the video sequence made up of image frames $i_j, \forall j \in [1, N]$. SGFB generates a spatio-temporal scene graph for every frame in the given video sequence. The scene graphs are encoded to formulate a spatio-temporal scene graph feature bank for the final task of action recognition. We describe the scene graph prediction and the scene graph feature bank components in more detail below. See Figure 5 for a high-level visualization of the model's forward pass.

### 4.1. Scene graph prediction

Previous research has proposed plenty of methods for predicting scene graphs on static images [48, 52, 77, 78, 85, 86]. We employ a state-of-the-art scene graph predictor as the first step of our method. Given a video sequence $v$, the scene graph predictor $SG$ generates all the objects and connects each object with their relationships with the actor in each frame, i.e. $SG : I \rightarrow G$. On each frame, the scene graph $G = (O, R)$ consists of a set of objects $O = \{o_1, o_2, \ldots\}$ that a person is interacting with and a set of relationships $R = \{\{r_{11}, r_{12}, \ldots\}, \{r_{21}, r_{22}, \ldots\}, \ldots\}$. Here $r_{pq}$ denotes the $q-$th relationship between the person with the object $o_p$. Note that there can be multiple relationships between the person and each object, including atten-

tion, spatial, and contact relationships. Besides the graph labels, the scene graph predictor $SG$ also outputs confidence scores for all predicted objects: $\{s_{o_1}, s_{o_2}, ...\}$ and relationships: $\{\{s_{r_{11}}, s_{r_{12}}, \dots\}, \{s_{r_{21}}, s_{r_{22}}, \dots\}, \dots\}$. We have experimented with various choices of $SG$ and benchmark their performance on Action Genome in Section 5.3.

### 4.2. Scene graph feature banks

After obtaining the scene graph $G$ on each frame, we formulate a feature vector $f$ by aggregating the information across all the scene graphs into a feature bank. Let's assume there are $|O|$ classes of objects and $|R|$ classes of relationships. In Action Genome, $|O| = 35$ and $|R| = 25$. We first construct a confidence matrix $C$ with dimension $|O| \times |R|$, where each entry corresponds to an object-relationship category pair. We compute every entry of this matrix using the scores output by the scene graph predictor $SG$. $C_{ij} = s_{o_i} \times s_{r_{ij}}$. Intuitively, $C_{ij}$ is a high value when $SG$ is confident that there is an object $o_i$ in the current frame and its relationship with the actor is $r_{ij}$. We flatten the confidence matrix as the feature vector $f$ for each image.

Formally, $F_{SG} = [f_1, f_2, ..., f_T]$ is a sequence of scene graph features extracted from a subsample of frames $i_1, i_2, ..., i_N$. We aggregate the features across the frames using methods similar to long-term feature banks [76], i.e. $F_{SG}$ are combined with 3D CNN features $S$ extracted from a short-term clip using feature bank operators (FBO), which can be instantiated as mean/max pooling or non-local blocks [73]. The 3D CNN embeds short-term information into $S$ while $F_{SG}$ provides contextual information, critical in modeling the dynamics of complex actions with long time span. The final aggregated feature is then used to predict action labels for the video.

## 5. Experiments

Action Genome's representation enables us to study few-shot action recognition by decomposing actions into temporally changing visual relationships between objects. It also allows us to benchmark whether understanding the decomposition helps improve performance in action recognition or scene graph prediction individually. To study these benefits afforded by Action Genome, we design three experiments: action recognition, few-shot action recognition, and finally, spatio-temporal scene graph prediction.

### 5.1. Action recognition on Charades

We expect that grounding the components that compose an action — the objects and their relationships — will improve our ability to predict which actions are occurring in a video sequence. So, we evaluate the utility of Action Genome's scene graphs on the task of action recognition.

**Problem formulation.** We specifically study multi-class action recognition on the Charades dataset [66]. The Charades dataset contains $9,848$ crowdsourced videos with a length of 30 seconds on average. At any frame, a person can perform multiple actions out of a nomenclature of 157 classes. The multi-classification task provides a video sequence as input and expects multiple action labels as output. We train our SGFB model to predict Charades action labels during test time and during training, provide SGFB with spatio-temporal scene graphs as additional supervision.

Table 3: Action recognition on Charades validation set in mAP (%). We outperform all existing methods when we simultaneously predict scene graphs while performing action recognition. We also find that utilizing ground truth scene graphs can significantly boost performance.

| Method | Backbone | Pre-train | mAP |
|---|---|---|---|
| `I3D + NL` [10, 73] | R101-I3D-NL | Kinetics-400 | 37.5 |
| `STRG` [74] | R101-I3D-NL | Kinetics-400 | 39.7 |
| `Timeception` [31] | R101 | Kinetics-400 | 41.1 |
| `SlowFast` [23] | R101 | Kinetics-400 | 42.1 |
| `SlowFast+NL` [23, 73] | R101-NL | Kinetics-400 | 42.5 |
| `LFB` [76] | R101-I3D-NL | Kinetics-400 | 42.5 |
| `SGFB` (ours) | R101-I3D-NL | Kinetics-400 | **44.3** |
| `SGFB Oracle` (ours) | R101-I3D-NL | Kinetics-400 | **60.3** |

**Baselines.** Previous work has proposed methods for multi-class action recognition and benchmarked on Charades. Recent state-of-the-art methods include applying I3D [10] and non-local blocks [73] as video feature extractors (`I3D+NL`), spatio-temporal region graphs (`STRG`) [74], Timeception convolutional layers (`Timeception`) [31], SlowFast networks (`SlowFast`) [23], and long-term feature banks (`LFB`) [76]. All the baseline methods are pre-trained on Kinetics-400 [38] and the input modality is RGB.

**Implementation details.** `SGFB` first predicts a scene graph on each frame, then constructs a spatio-temporal scene graph feature bank for action recognition. We use Faster R-CNN [62] with ResNet-101 [29] as the backbone for region proposals and object detection. We leverage RelDN [86] to predict the visual relationships. Scene graph prediction is trained on Action Genome, where we follow the same train/val splits of videos as the Charades dataset. Action recognition uses the same video feature extractor, hyper-parameters, and solver schedulers as long-term feature banks (`LFB`) [76] for a fair comparison.

**Results.** We report performance of all models using mean average precision (mAP) on Charades validation set in Table 3. By replacing the feature banks with spatio-temporal scene graph features, we outperform the state-of-the-art LFB by $1.8\%$ mAP. Our features are smaller in size ($35 \times 25 = 875$ in SGFB versus 2048 in LFB) but concisely capture the more information for recognizing actions.

We also find that improving object detectors designed for videos can further improve action recognition results. To quantitatively demonstrate the potential of better

Table 4: Few-shot experiments. With the ability of compositional action understanding, our SGFB demonstrates better generalizability than LFB. The SGFB oracle shows the great potential of how much the scene graph representation could benefit action recognition.

| | 1-shot | 5-shot | 10-shot |
|---|---|---|---|
| LFB [76] | 28.3 | 36.3 | 39.6 |
| SGFB (ours) | **28.8** | **37.9** | **42.7** |
| SGFB oracle (ours) | **30.4** | **40.2** | **50.5** |

scene graphs on action recognition, we designed an SGFB Oracle setup. The SGFB Oracle assumes that a perfect scene graph prediction method is available. The spatio-temporal scene graph feature bank therefore, directly encodes a feature vector from ground truth objects and visual relationships for the annotated frames. Feeding such feature banks into the SGFB model, we observe a significant improvement on action recognition: 16% increase on mAP. Such a boost in performance shows the potential of Action Genome and compositional action understanding when video-based scene graph models are utilized to improve scene graph prediction. It is important to note that the performance by SGFB Oracle is not an upper bound on performance since we only utilize ground truth scene graphs for the few frames that have ground truth annotations.

### 5.2. Few-shot action recognition

Intuitively, predicting actions should be easier from a symbolic embedding of scene graphs than from pixels. When trained with very few examples, compositional action understanding with additional knowledge of scene graphs should outperform methods that treat actions as monolithic concept. We showcase the capability and potential of spatio-temporal scene graphs to generalize to rare actions.

**Problem formulation.** In our few-shot action recognition experiments on Charades, we split the 157 action classes into a base set of 137 classes and a novel set of 20 classes. We first train a backbone feature extractor (R101-I3D-NL) on all video examples of the base classes, which is shared by the baseline LFB, our SGFB, and SGFB oracle. Next, we train each model with only $k$ examples from each novel class, where $k = 1, 5, 10$, for 50 epochs. Finally, we evaluate the trained models on all examples of novel classes in the Charades validation set.

**Results.** We report few-shot experiment performance in Table 4. SGFB achieves better performance than LFB on all $1, 5, 10$-shot experiments. Furthermore, if with ground truth scene graphs, SGFB Oracle shows a 10.9% 10-shot mAP improvement. We visualize the comparison between SGFB and LFB in Figure 6. With the knowledge of spatio-temporal scene graphs, SGFB better captures action concepts involving the dynamics of objects and relationships.

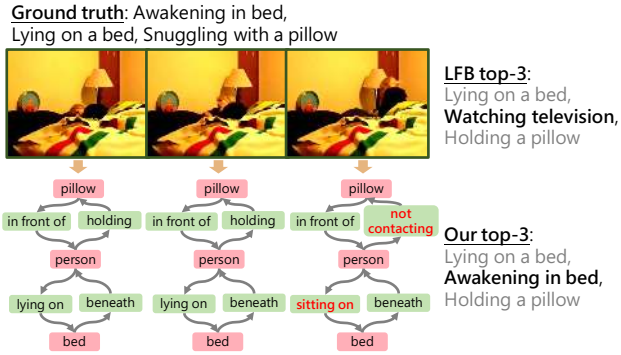


Figure 6: Qualitative results of 10-shot experiments. We compare the predictions of our SGFB against LFB [76]. Since SGFB uses scene graph knowledge and explicitly captures the dynamics of human-object relationships, it easily learns the concept of "awakening in bed" even when only trained with 10 examples of this label. Also, since SGFB is trained to detect and ground objects, it avoids misclassifying objects, such as television, which then results in more robust action recognition.

### 5.3. Spatio-temporal scene graph prediction

Progress in image-based scene graph prediction has cascaded to improvements across multiple Computer Vision tasks, including image captioning [2], image retrieval [36, 64], visual question answering [35], relationship modeling [41] and image generation [34]. In order to promote similar progress in video-based tasks, we introduce the complementary of spatio-temporal scene graph prediction. Unlike image-based scene graph prediction, which only has a single image as input, this task expects a video as input and therefore, can utilize temporal information from neighboring frames to strength its predictions. In this section, we define the task, its evaluation metrics and report benchmarked results from numerous recently proposed image-based scene graph models applied to this new task.

**Problem formulation.** The task expects as input a video sequence $v = \{i_1, i_2, \dots i_n\}$ where $i_j \forall j \in [1, n]$ represents image frames from the video. The task requires the model to generate a spatio-temporal scene graph $G = (O, R)$ per frame. $o_k \in O$ is represented as objects with category labels and bounding box locations. $r_{j,kl} \in R$ represents the relationships between objects $o_i$ and $o_j$.

**Evaluation metrics.** We borrow the three standard evaluation modes for image-based scene graph prediction [52]: (i) scene graph detection (SGDET) which expects input images and predicts bounding box locations, object categories, and predicate labels, (ii) scene graph classification (SGCLS) which expects ground truth boxes and predicts object categories and predicate labels, and (iii) predicate classification (PREDCLS), which expects ground truth bounding boxes

Table 5: We evaluate recently proposed image-based scene graph prediction models and provide a benchmark for the new task of spatio-temporal scene graph prediction. We find that there is significant room for improvement, especially since these existing methods were designed to be conditioned on a single frame and do not consider the entire video sequence as a whole.

| Method | PredCls | | | | SGCls | | | | SGGen | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | image | | video | | image | | video | | image | | video | |
| | R@20 | R@50 | R@20 | R@50 | R@20 | R@50 | R@20 | R@50 | R@20 | R@50 | R@20 | R@50 |
| VRD [52] | 14.75 | 14.85 | 14.51 | 14.60 | 13.65 | 14.69 | 13.41 | 14.44 | 10.28 | 10.94 | 10.04 | 10.70 |
| Freq Prior [85] | 32.70 | 32.84 | 32.25 | 32.37 | 31.52 | 32.78 | 31.08 | 32.32 | 24.03 | 24.87 | 23.49 | 24.31 |
| IMP [77] | 35.15 | 35.56 | 34.50 | 34.86 | 31.73 | 34.85 | 31.09 | 34.16 | 23.88 | 25.52 | 23.23 | 24.82 |
| MSDN [48] | 35.27 | 35.64 | 34.61 | 34.93 | 31.89 | 34.98 | 31.28 | 34.28 | 24.00 | 25.64 | 23.39 | 24.95 |
| Graph R-CNN [78] | 35.36 | 35.74 | 34.80 | 35.12 | 31.94 | 35.07 | 31.43 | 34.46 | 24.12 | 25.77 | 23.59 | 25.15 |
| RelDN [86] | **35.89** | **36.09** | **35.36** | **35.51** | **33.47** | **35.84** | **32.96** | **35.27** | **25.00** | **26.21** | **24.45** | **25.63** |

and object categories to predict predicate labels. We refer the reader to the paper that introduced these tasks for more details [52]. We adapt these metrics for video, where the per-frame measurements are first averaged in each video as the measurement of the video, then we average video results as the final result for the test set.

**Baselines.** We benchmark the following image-based scene graph models for the spatio-temporal scene graph prediction task: VRD's visual module (VRD) [52], neural motif's frequency prior (Freq-prior) [85], iterative message passing (IMP) [77], multi-level scene description network (MSDN) [48], graph R-CNN (Graph R-CNN) [78], and relationship detection network (RelDN) [86].

**Results.** To our surprise, we find that IMP, which was one of the earliest scene graph prediction models actually outperforms numerous more recently proposed methods. The most recently proposed scene graph model, RelDN marginally outperforms IMP, suggesting that modeling similarlities between object and relationship classes improve performance in our task as well. The small gap in performance between the task of PredCls and SGCls suggests that these models suffer from not being able to accurately detect the objects in the video frames. Improving object detectors designed specifically for videos could improve performance. The models were trained only using Action Genome's data and not finetuned on Visual Genome [43], which contains image-based scene graphs, or on ActivityNet Captions [42], which contains dense captioning of actions in videos with natural language paragraphs. We expect that finetuning models with such datasets would result in further improvements.

## 6. Future work

With the rich hierarchy of events, Action Genome not only enables research on spatio-temporal scene graph prediction and compositional action recognition, but also promises various research directions. We hope future work will develop methods for the following:

**Spatio-temporal action localization.** The majority of spatio-temporal action localization methods [24, 25, 33, 68] focus on localizing the person performing the action but ignore the objects, which are also involved in the action, that the person interacts with. Action Genome can enable research on localization of both actors and objects, formulating a more comprehensive grounded action localization task. Furthermore, other variants of this task can also be explored; for example, a weakly-supervised localization task where a model is trained with only action labels but tasked with localizing the actors and objects.

**Explainable action models.** Explainable visual models is an emerging field of research. Amongst numerous techniques, saliency prediction has emerged as a key mechanism to interpret machine learning models [54, 65, 67]. Action Genome provides frame-level labels of attention in the form of objects that a the person performing the action is either looking at or interacting with. These labels can be used to further train explainable models.

**Video generation from spatio-temporal scene graphs.** Recent studies have explored image generation from scene graphs [3, 34]. Similarly, with a structured video representation, Action Genome enables research on video generation from spatio-temporal scene graphs.

## 7. Conclusion

We introduce Action Genome, a representation that decomposes actions into spatio-temporal scene graphs. Scene graphs explain how objects and their relationships change as an action occurs. We demonstrated the utility of Action Genome by collecting a large dataset of spatio-temporal scene graphs and used it to improve state of the art results for action recognition as well as few-shot action recognition. Finally, we benchmarked results for the new task of spatio-temporal scene graph prediction. We hope that Action Genome will inspire a new line of research in more decomposable and generalizable video understanding.

**Acknowledgement.** We would like to thank Panasonic for their support.

## References


[1] Tim Althoff, Hyun Oh Song, and Trevor Darrell. Detection bank: an object detection based video representation for multimedia event recognition. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1065–1068. ACM, 2012. 5

[2] 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. 1, 3, 7

[3] Oron Ashual and Lior Wolf. Specifying object attributes and relations in interactive scene generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4561–4569, 2019. 8

[4] Fabien Baradel, Natalia Neverova, Christian Wolf, Julien Mille, and Greg Mori. Object level visual reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 105–121, 2018. 3

[5] Roger G Barker and Herbert F Wright. One boy's day; a specimen record of behavior. 1951. 1, 2

[6] Roger G Barker and Herbert F Wright. Midwest and its children: The psychological ecology of an american town. 1955. 1, 2

[7] Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. *arXiv preprint arXiv:1907.09021*, 2019. 3

[8] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 1, 2, 3

[9] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 2, 3

[10] 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. 1, 3, 6

[11] Roberto Casati and A Varzi. Events, volume 15 of the international research library of philosophy, 1996. 2

[12] Vincent S Chen, Paroma Varma, Ranjay Krishna, Michael Bernstein, Christopher Re, and Li Fei-Fei. Scene graph prediction with limited labels. *arXiv preprint arXiv:1904.11622*, 2019. 3

[13] Aron Culotta and Jeffrey Sorensen. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd annual meeting on association for computational linguistics*, page 423. Association for Computational Linguistics, 2004. 3

[14] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3298–3308. IEEE, 2017. 3

[15] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 2, 3

[16] Chaitanya Desai, Deva Ramanan, and Charless C Fowlkes. Discriminative models for multi-class object layout. *International journal of computer vision*, 95(1):1–12, 2011. 3

[17] Apoorva Dornadula, Austin Narcomey, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationships as functions: Enabling few-shot scene graph prediction. *arXiv preprint arXiv:1906.04876*, 2019. 3

[18] Sai Kumar Dwivedi, Vikram Gupta, Rahul Mitra, Shuaib Ahmed, and Arjun Jain. Protogan: Towards few shot learning for action recognition. *arXiv preprint arXiv:1909.07945*, 2019. 3

[19] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Daps: Deep action proposals for action understanding. In *European Conference on Computer Vision*, pages 768–784. Springer, 2016. 2

[20] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, 2009. 3

[21] Li Fei-Fei, Rob Fergus, and Pietro Perona. A bayesian approach to unsupervised one-shot learning of object categories. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1134–1141. IEEE, 2003. 3

[22] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006. 3

[23] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6202–6211, 2019. 1, 3, 6

[24] Rohit Girdhar, João Carreira, Carl Doersch, and Andrew Zisserman. A better baseline for ava. *arXiv preprint arXiv:1807.10066*, 2018. 8

[25] 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. 3, 8

[26] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 2, 3

[27] Zhou GuoDong, Su Jian, Zhang Jie, and Zhang Min. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 427–434. Association for Computational Linguistics, 2005. 3

[28] Bridgette M Hard, Barbara Tversky, and David S Lang. Making sense of abstract events: Building event schemas. *Memory & cognition*, 34(6):1221–1235, 2006. 2

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceed-*

*ings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[30] Roei Herzig, Moshiko Raboh, Gal Chechik, Jonathan Berant, and Amir Globerson. Mapping images to scene graphs with permutation-invariant structured prediction. In *Advances in Neural Information Processing Systems*, pages 7211–7221, 2018. 3

[31] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Timeception for complex action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 254–263, 2019. 1, 3, 6

[32] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 5308–5317, 2016. 3

[33] Jianwen Jiang, Yu Cao, Lin Song, Shiwei Zhang4 Yunkai Li, Ziyao Xu, Qian Wu, Chuang Gan, Chi Zhang, and Gang Yu. Human centric spatio-temporal action localization. In *ActivityNet Workshop on CVPR*, 2018. 8

[34] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. *arXiv preprint arXiv:1804.01622*, 2018. 1, 3, 7, 8

[35] 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. 1, 3, 7

[36] 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. 1, 3, 7

[37] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 2

[38] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 6

[39] Orit Kliper-Gross, Tal Hassner, and Lior Wolf. One shot similarity metric learning for action recognition. In *International Workshop on Similarity-Based Pattern Recognition*, pages 31–45. Springer, 2011. 3

[40] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011. 3

[41] Ranjay Krishna, Ines Chami, Michael Bernstein, and Li Fei-Fei. Referring relationships. In *Computer Vision and Pattern Recognition*, 2018. 1, 3, 7

[42] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 8

[43] 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, 3, 8

[44] Christopher A Kurby and Jeffrey M Zacks. Segmentation in the perception and memory of events. *Trends in cognitive sciences*, 12(2):72–79, 2008. 1, 2, 3

[45] Li-Jia Li, Hao Su, Li Fei-Fei, and Eric P Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in neural information processing systems*, pages 1378–1386, 2010. 5

[46] Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'Ou Tang. Vip-cnn: Visual phrase guided convolutional neural network. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 7244–7253. IEEE, 2017. 3

[47] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: an efficient subgraph-based framework for scene graph generation. In *European Conference on Computer Vision*, pages 346–363. Springer, 2018. 3

[48] 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. 3, 5, 8

[49] Xiaodan Liang, Lisa Lee, and Eric P Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4408–4417. IEEE, 2017. 3

[50] Ivan Lillo, Alvaro Soto, and Juan Carlos Niebles. Discriminative hierarchical modeling of spatio-temporally composable human activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 812–819, 2014. 1, 2

[51] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7083–7093, 2019. 3

[52] 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. 5, 7, 8

[53] Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, and Hans Peter Graf. Attend and interact: Higher-order object interactions for video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6790–6800, 2018. 3

[54] Aravindh Mahendran and Andrea Vedaldi. Salient deconvolutional networks. In *European Conference on Computer Vision*, pages 120–135. Springer, 2016. 8

[55] Albert Michotte. *The perception of causality*. Routledge, 1963. 1, 2

[56] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 2

[57] George A Miller and Philip N Johnson-Laird. *Language and perception*. Belknap Press, 1976. 2

[58] Ashish Mishra, Vinay Kumar Verma, M Shiva Krishna Reddy, S Arulkumar, Piyush Rai, and Anurag Mittal. A generative approach to zero-shot and few-shot action recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 372–380. IEEE, 2018. 3

[59] Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. In *Advances in Neural Information Processing Systems*, pages 2168–2177, 2017. 3

[60] Darren Newtson. Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, 28(1):28, 1973. 2

[61] Devi Parikh and Kristen Grauman. Relative attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 503–510. IEEE, 2011. 3

[62] 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. 6

[63] Jeremy R Reynolds, Jeffrey M Zacks, and Todd S Braver. A computational model of event segmentation from perceptual prediction. *Cognitive science*, 31(4):613–643, 2007. 2, 3

[64] 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. 1, 3, 7

[65] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 8

[66] 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. 1, 2, 3, 4, 6

[67] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 8

[68] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-centric relation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 318–334, 2018. 8

[69] 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. 1, 3

[70] Zhuowen Tu and Xiang Bai. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(10):1744–1757, 2010. 3

[71] Gül Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1510–1517, 2017. 2

[72] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 1, 3

[73] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 6

[74] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 399–417, 2018. 3, 6

[75] Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. Human action localization with sparse spatial supervision. *arXiv preprint arXiv:1605.05197*, 2016. 2, 3

[76] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019. 2, 3, 5, 6, 7

[77] 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. 3, 5, 8

[78] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. *arXiv preprint arXiv:1808.00191*, 2018. 3, 5, 8

[79] Benjamin Yao, Xiong Yang, and Song-Chun Zhu. Introduction to a large-scale general purpose ground truth database: methodology, annotation tool and benchmarks. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 169–183. Springer, 2007. 2

[80] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 126(2-4):375–389, 2018. 2

[81] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2678–2687, 2016. 2

[82] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015. 2

[83] Jeffrey M Zacks, Todd S Braver, Margaret A Sheridan, David I Donaldson, Abraham Z Snyder, John M Ollinger, Randy L Buckner, and Marcus E Raichle. Human brain activity time-locked to perceptual event boundaries. *Nature neuroscience*, 4(6):651, 2001. 2

[84] Jeffrey M Zacks, Barbara Tversky, and Gowri Iyer. Perceiving, remembering, and communicating structure in events.

*Journal of experimental psychology: General*, 130(1):29, 2001. 1, 2, 3

[85] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. *arXiv preprint arXiv:1711.06640*, 2017. 3, 5, 8

[86] 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. 5, 6, 8

[87] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8668–8678, 2019. 1, 2, 3

[88] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 3

[89] Guodong Zhou, Min Zhang, DongHong Ji, and Qiaoming Zhu. Tree kernel-based relation extraction with context-sensitive structured parse tree information. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007. 3

[90] Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 751–766, 2018. 3

[91] 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. 2