Something-Else: Compositional Action Recognition with Spatial-Temporal Interaction Networks

Joanna Materzynska
University of Oxford, TwentyBN

Tete Xiao
UC Berkeley

Roei Herzig
Tel Aviv University

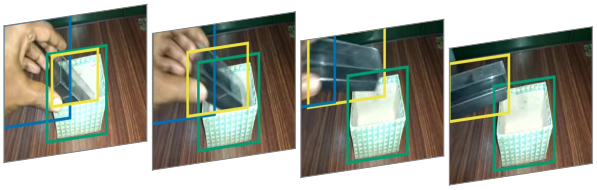
Huijuan Xu*
UC Berkeley

Xiaolong Wang*
UC Berkeley

Trevor Darrell*
UC Berkeley

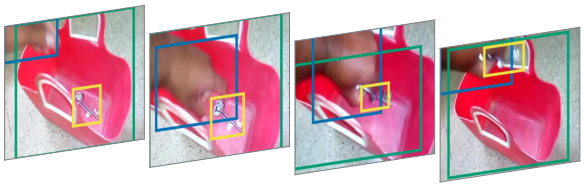
Abstract

Human action is naturally compositional: humans can easily recognize and perform actions with objects that are different from those used in training demonstrations. In this paper, we study the compositionality of action by looking into the dynamics of subject-object interactions. We propose a novel model which can explicitly reason about the geometric relations between constituent objects and an agent performing an action. To train our model, we collect dense object box annotations on the Something-Something dataset. We propose a novel compositional action recognition task where the training combinations of verbs and nouns do not overlap with the test set. The novel aspects of our model are applicable to activities with prominent object interaction dynamics and to objects which can be tracked using state-of-the-art approaches; for activities without clearly defined spatial object-agent interactions, we rely on baseline scene-level spatio-temporal representations. We show the effectiveness of our approach not only on the proposed compositional action recognition task, but also in a few-shot compositional setting which requires the model to generalize across both object appearance and action category.[1]



STIN: *Taking smth out of smth* I3D: *Taking smth out of smth*

(a) Seen verb and object combination



STIN: *Taking smth out of smth* I3D: *Poking a hole into sthm soft*

(b) Unseen verb and object combination

Figure 1. Two example videos of an action class "taking something out of something": the activity defines the relative change in object and agent (hand) positions over time. Most current methods (I3D-based) over-rely on object appearance. While it works well on seen verb and object combination in (a), it cannot generalize to unseen combinations in (b). Our Spatial-Temporal Interaction Networks (STIN) is designed for generalizing action recognition regardless of the object appearance in the training set. (Correct predictions are in green, incorrect in red.)

1. Introduction

Let's look at the simple action of "taking something out of something" in Figure 1. Even though these two videos show human hands interacting with different objects, we recognize that they are the same action based on changes in the relative positions of the objects and hands involved in the activity. Further, we can easily recognize the action even when it is presented with previously unseen objects and tools. We ask, do current machine learning algorithms have the capability to generalize across different combinations of verbs and nouns?

We investigate actions represented by the changes in geometric arrangements between subjects (agents) and objects. We propose a compositional action recognition setting in which we decompose each action into a combination of a verb, a subject, and one or more objects. Instead of the traditional setting where training and testing splits include the same combinations of verbs and nouns, we train

*Equal advising

[1]Project page: https://joaanna.github.io/something_else/.

and test our model on the same set of verbs (actions) but combine them with different object categories, so that tested verb and object combinations have never been seen during training time (Figure 1 (b)).

This problem turns out to be very challenging for heretofore state-of-the-art action recognition models. Computer vision researchers have developed deep networks with temporal connections for action recognition by using Recurrent Neural Networks with 2D Convolutions [76, 11] and 3D ConvNets [7, 74, 61, 63]. However, both types of models have difficulty in this setting; our results suggest that they cannot fully capture the compositionality of action and objects. These approaches focus on extracting features for the whole scene and do not explicitly recognize objects as individual entities; scene-level convolutional operators may rely more on spatial appearance rather than temporal transformations or geometric relations, since the former alone are often highly predictive of the action class [54, 3].

Recently, researchers have investigated building spatial-temporal graph representations of videos [67, 68, 9, 25] leveraging recently proposed graph neural networks [38]. These methods take dense object proposals as graph nodes and learn the relations between them. While this certainly opens a door for bringing relational reasoning in video understanding, the improvement over the 3D ConvNet baselines is not very significant. Generally, these methods have employed non-specific object graphs based on a large set of object proposals in each frame, rather than sparse semantically grounded graphs which model the specific interaction of an agent and constituent objects in an action.

In this paper, we propose a model based on a sparse and semantically-rich object graph learned for each action. We train our model with accurately localized object boxes in the demonstrated action. Our model learns explicit relations between subjects and objects; these turn out to be the key for successful compositional action recognition. We leverage state-of-the-art object detectors to accurately locate the subject (agent) and constituent objects in the videos, perform multi-object tracking on them and form multiple tracklets for boxes belonging to the same instance. As shown in Figure 1, we localize the hand, and the objects manipulated by the hand. We track the objects over time and the objects belonged to the same instance are illustrated by the boxes with the same color.

Our Spatial-Temporal Interaction Network (STIN) reasons on candidate sparse graphs found from these detection and tracking results. Our model takes the locations and shapes of objects and subject in each frame as inputs. It first performs spatial interaction reasoning on them by propagating the information among the subjects and objects, then we perform temporal interaction reasoning over the boxes along the same tracklet, which encodes the transformation of objects and the relation between subjects and objects in time. Finally, we compose the trajectories for the agent and the objects together to understand the action. Our model is designed for activities which have prominent interaction dynamics between a subject or agent (*e.g.*, hand) and constituent objects; for activities where no such dynamics are clearly discernible with current detectors (*e.g.*, pouring water, crushing paper), our model falls back to leverage baseline spatio-temporal scene representations.

We introduce the Something-Else task, which extends the Something-Something dataset [20] with new annotations and a new compositional split. In our compositional split, methods are required to recognize an action when performed with unseen objects, i.e., objects which do not appear together with this action at training time. Thus methods are trained on "Something", but are tested on their ability to generalize to "Something-Else". Each action category in this dataset is described as a phrase composed with the same verb and different nouns. We reorganize the dataset for compositional action recognition and model the dynamics of inter-object geometric configurations across time per action. We investigate compositional action recognition tasks in both a standard setting (where training and testing are with the same categories) and a few-shot setting (where novel categories are introduced with only a few examples). To support these two tasks, we collect and will release annotations on object bounding boxes for each video frame. Surprisingly, we observe even with only low dimensional coordinate inputs, our model can show comparable results and improves the appearance-based models in few-shot setting by a significant margin.

Our contributions include: (i) A Spatial-Temporal Interaction Network which explicitly models the changes of geometric configurations between agents and objects; (ii) Two new compositional tasks for testing model generalizability and dense object bounding box annotations in videos; (iii) Substantial performance gain over appearance-based model on compositional action recognition.

2. Related Work

Action recognition is of central importance in computer vision. Over the past few years, researchers have been collecting larger-scale datasets including Jester [44], UCF101 [59], Charades [56], Sports1M [35] and Kinetics [37]. Boosted by the scale of data, modern deep learning approaches, including two-stream ConvNets [57, 66], Recurrent Neural Networks [76, 12, 47, 5] and 3D ConvNets [31, 7, 14, 73, 62, 63, 13], have shown encouraging results on these datasets. However, a recent study in [77] indicates that most of the current models trained with the above-mentioned datasets are not focusing on temporal reasoning but the appearance of the frames: Reversing the order of the video frames at test time will lead to almost the same classification result. In light of this problem, the Something-Something dataset [20] is introduced to recog-

nize action independent of the object appearance. To push this direction forward, we propose the compositional action recognition task for this dataset and provide object bounding box annotations.

The idea of compositionality in computer vision originates from Hoffman's research on Parts of Recognition [26]. Following this work, models with pictorial structures have been widely studied in traditional computer vision [15, 78, 29]. For example, Felzenszwalb *et al*. [15] proposes a deformable part-based model that organizes a set of part classifiers in a deformable manner for object detection. The idea of composing visual primitives and concepts has also been brought back in the deep learning community recently [64, 45, 36, 1, 32, 28]. For example, Misra *et al*. [45] propose a method to compose classifiers of known visual concepts and apply this model to recognize objects with unseen combinations of concepts. Motivated by this work, we propose to explicitly compose the subjects and objects in a video and reason about the relationships between them to recognize the action with unseen combinations of verbs and nouns.

The study of visual relationships has a long history in computer vision [23, 75, 51] and early work investigated combining object and motion features for action recognition [52, 46]. Recent works have shown relational reasoning with deep networks on images [19, 27, 55, 33]. For example, Gkioxari *et al*. [19] proposes to accurately detect the relations between the objects together with state-of-the-art object detectors. The idea of relational reasoning has also been extended in video understanding [67, 68, 70, 25, 60, 18, 2, 69, 30]. For instance, Wang *et al*. [68] apply a space-time region graph to improve action classification in cluttered scenes. Instead of only relying on dense "objectness" region proposals, Wu *et al*. [70] further extend this graph model with accurate human detection and reasoning over a longer time range. Motivated by these works, we build our spatial-temporal interaction network to reason about the relations between subjects and objects based on accurate detection and tracking results. Our work is also related to the Visual Interaction Network [69], which models the physical interactions between objects in a simulated environment.

To further illustrate the generalizability of our approach, we also apply our model in a few-shot setting. Few-shot image recognition has become a popular research topic in recent years [16, 58, 65, 8, 17, 48]. Chen *et al*. [8] has re-examined recent approaches in few-shot learning and found a simple baseline model which is very competitive compared to meta-learning approaches [16, 40, 53]. Researchers have also investigated few-shot learning in videos [22, 6]. Guo *et al*. [22] propose to perform KNN on object graph representations for few-shot 3D action recognition. We adopt our spatial-temporal interaction network for few-shot video classification, by using the same learning scheme as the simple baseline mentioned in [8].

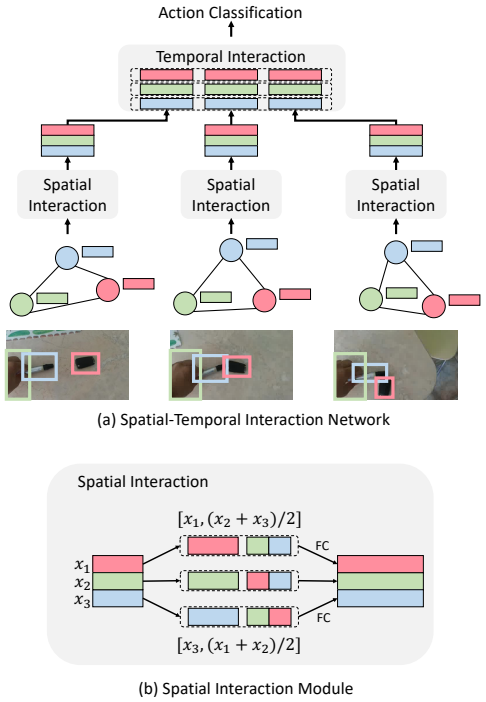


Figure 2. (a) The Spatial-Temporal Interaction Network (STIN): our model operates on object-centric features and performs spatial interaction reasoning for individual frames and temporal interaction reasoning to obtain a classification decision. (Different colors represent different objects in this figure.) (b) The Spatial Interaction Module: Given a set of object features in one frame, we aggregate them together with the information about their relative position by applying Eq. 1 to update each object feature.

3. Spatial-Temporal Interaction Networks

We present Spatial-Temporal Interaction Networks (STIN) for compositional action recognition. Our model utilizes a generic detector and tracker to build object-graph representations that explicitly include hand and constituent object nodes. We perform spatial-temporal reasoning among these bounding boxes to understand how the relations between subjects and objects change over time for a given action (Figure 2). By explicitly modeling the transformation of object geometric relations in a video, our model can effectively generalize to videos with unseen combinations of verbs and nouns as demonstrated in Figure 3.

3.1. Object-centric Representation

Given a video with T frames, we first perform object detection on these video frames, using a detector which detects hands and generic candidate constituent objects. The object detector is trained on the set of all objects in the train split of the dataset as one class, and all hands in the training data as a second class. Assume that we have detected

N instances including the hands and the objects manipulated by the hands in the scene, we then perform multi-object tracking to find correspondences between boxes in different video frames. We extract two types of feature representation for each box: (a) bounding box coordinates; and (b) an object identity feature. Both of these features are designed for compositional generalization and avoiding object appearance bias.

Bounding box coordinates. One way to represent an object and its movement is to use its location and shape. We use the center coordinate of each object along with its height and width as a quadruple, and forward it to a Multi-Layer Perceptron (MLP), yielding a d-dimensional feature. Surprisingly, this simple representation alone turns out to be highly effective in action recognition.

Object identity embedding. In addition to the object coordinate feature, we also use a learnable d-dimensional embedding to represent the identities of objects and subjects. We define three types of embedding: (i) *subject* (or *agent*) embedding, *i.e.*, representing hands in an action; (ii) *object* embedding, *i.e.*, representing the objects involved in the action; (iii) *null* embedding, *i.e.*, representing dummy boxes irrelevant to the action. The three embeddings are initialized from an independent multivariate normal distribution. The identity embedding can be concatenated together with box coordinate features as the input to our model. Since the identity (category) of the instances is predicted by the object detector, we can combine coordinate features with embedding features accordingly. We note that these embeddings do not depend on the appearance of input videos.

We find that combining the box coordinate feature with the identity feature significantly improves the performance of our model. Since we are using a *general object* embedding for all kinds of objects, this helps the model to generalize across different combinations of verbs and nouns in a compositional action recognition setting.

Robustness to Unstable Detection. In cases where object detector is not reliable, where the number of detected objects is larger than a fix number N, we can perform object configuration search during inference. Each time we randomly sample N object tracklets and forward them to our model. We perform classification based on the most confident configuration which has the highest score. However, in our current experiments, we can already achieve significant improvement without this process.

3.2. Spatial-temporal interaction reasoning

Given T video frames and N objects per frame, we denote the set of object features as $X = (x_1^1, ..., x_N^1, x_1^2, ..., x_N^2, ..., x_N^T)$, where x_i^t represents the feature of object i in frame t. Our goal is to perform spatial-temporal reasoning in X for action recognition. As illustrated in Figure 2(a), we first perform spatial interaction reasoning on objects in each frame, then we connect these features together with temporal interaction reasoning.

Spatial interaction module. We perform spatial interaction reasoning among the N objects in each frame. For each object x_i^t, we first aggregate the features from the other $N-1$ objects by averaging them, then we concatenate the aggregated feature with x_i^t. This process can be represented as,

$$f(x_i^t) = \text{ReLU}(W_f^T[x_i^t, \frac{1}{N-1}\sum_{j \neq i} x_j^t]), \quad (1)$$

where $[,]$ denotes concatenation of two features in the channel dimension and W_f^T is learnable weights implemented by a fully connected layer. We visualize this process in Figure 2(b) in the case of $N = 3$.

Temporal interaction module. Given the aggregated feature of objects in each frame, we perform temporal reasoning on top of the features. As tracklets are formed and obtained previously, we can directly link objects of the same instance across time. Given objects in the same tracklet, we compute the feature of the tracklet as $g(x_i^1, ..., x_i^T)$: We first concatenate the object features, then forward the combined feature to another MLP network. Given a set of temporal interaction results, we aggregate them together for action recognition as,

$$p(X) = W_p^T h(\{g(x_i^1, ..., x_i^T)\}_{i=1}^N), \quad (2)$$

where h is a function combining and aggregating the information of tracklets. In this study, we experiment with two different approaches to combine tracklets: (i) Design h as a simple averaging function to prove the effectiveness of our spatial-temporal interaction reasoning. (ii) Utilize non-local block [67] as the function h. The non-local block encodes the pairwise relationships between every two trajectory features before averaging them. In our implementation, we adopt three non-local blocks succeeded by convolutional kernels. We use W_p as our final classifier with cross-entropy loss.

Combining video appearance representation. Besides explicitly modeling the transformation of relationships of subjects and objects, our spatial-temporal interaction model can be easily combined with any video-level appearance representation. The presence of appearance features helps especially the action classes without prominent inter-object dynamics. To achieve this, we first forward the video frames to a 3D ConvNet. We follow the network backbone applied in [68], which takes T frames as input and extracts a spatial-temporal feature representation. We perform average pooling across space and time on this feature representation, yielding a d-dimensional feature. Video appearance representations are concatenated with object representations $h(\{g(x_i^1, ..., x_i^T)\}_{i=1}^N)$, before fed into the classifier.

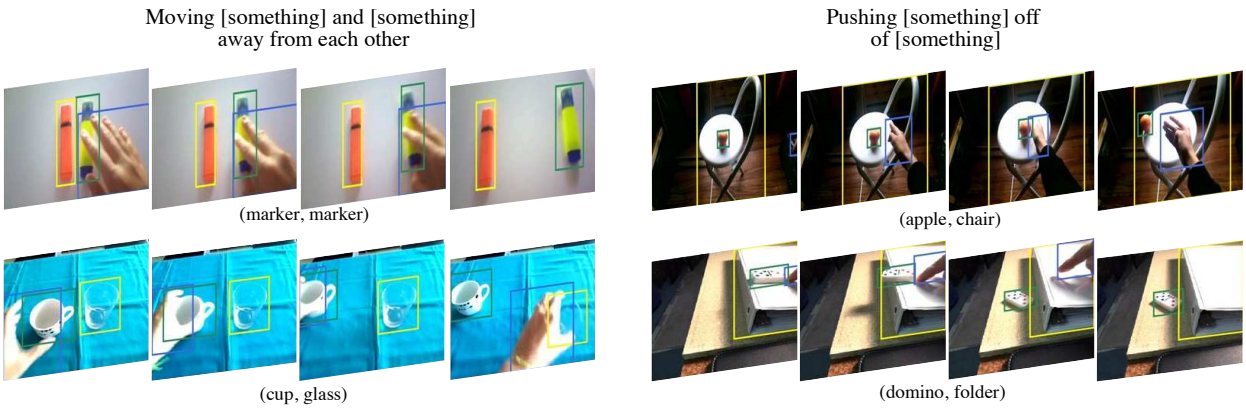


Figure 3. Annotated examples of the Something-Something V2 dataset. Understanding the action from the visual appearance of the entire scene is challenging because we can perform the same action using arbitrary objects, however, observing the relative change of the location and positioning of the object and hands in the scene concisely captures the interaction.

4. The Something-Else Task

To present the idea of compositional action recognition, we adopt the Something-Something V2 dataset [20] and create new annotations and splits within it. We name the action recognition on the new splits as the "Something-Else task".

The Something-Something V2 dataset contains 174 categories of common human-object interactions. Collected via Amazon Mechanical Turk in a crowd-sourced manner, the protocol allows turkers to pick an action category (*verb*), perform and upload a video accordingly with arbitrary objects (*noun*). The lack of constraints in choosing the objects naturally results in large variability in the dataset. There are $12,554$ different object descriptions in total. The original split does not consider the distribution of the objects in the training and the testing set, instead, it asserts that the videos recorded by the same person are in either training or testing set but both. While this setting reduces the environment and individual bias, it ignores the fact that the combination of verbs and nouns presented in the testing set may have been encountered in the training stage. The high performance obtained in this setting might indicate that models have learned the actions coupled by typical objects occurring, yet does not reflect the generalization capacity of models to actions with novel objects.

Compositional Action Recognition. In contrast to randomly assigning videos into training or testing sets, we present a compositional action recognition task. In our setting, the combinations of a verb (action) and nouns in the training set do not exist in the testing set. We define a subset of *frequent object categories* as those appearing in more than 100 videos in the dataset. We split the *frequent object categories* into two disjoint groups, $\mathcal{A}$ and $\mathcal{B}$. Besides objects, action categories are divided into two groups 1 and 2 as well. In [20] these categories are organized hierarchically, *e.g.*, *"moving something up"* and *"moving something down"* belong to the same super-class. We randomly assign each action category into one of two groups, and at the same time enforce that the actions belonging to the same super-class are assigned into the same group.

Task Split	# Classes	Training	Validation
Original	174	168,913	24,777
Compositional	174	54,919	57,876
FS-Base	88	112,397	12,467
FS-Novel 5-S	86	430	49,822
FS-Novel 10-S	86	860	43,954

Table 1. **Comparison and statistics of various tasks** on the Something-Something V2. FS: few-shot; n-S: n-shot.

Given the splits of groups, we combine action group 1 with object group $\mathcal{A}$, and action group 2 with object group $\mathcal{B}$, to form the training set, termed as $1\mathcal{A}+2\mathcal{B}$. The validation set is built by flipping the combination into $1\mathcal{B}+2\mathcal{A}$. Different combinations of verbs and nouns are thus divided into training *or* testing splits in this way. The statistics of the training and the validation sets under the compositional setting are shown in the second row of Table 1.

Few-shot Compositional Action Recognition. The compositional split challenges the network to generalize over object appearance. We further consider a few-shot dataset split setting indicating how well a trained action recognition model can generalize to novel action categories with only a few training examples. We assign the action classes in the Something-Something V2 dataset into a *base* split and a *novel* split, yielding 88 classes in the base set and 86 classes in the novel set. We randomly allocate 10% of the videos from the base set to form a validation set and the rest of the videos as the base training set. We then randomly select k examples for each category in the novel set whose labels are present in the training stage, and the remaining videos from the novel set are designated as the validation set. We ensure that the object categories in k-shot training videos do not appear in the novel validation set. In this way, our few-shot setting additionally challenges models to generalize over object appearance. We term this task

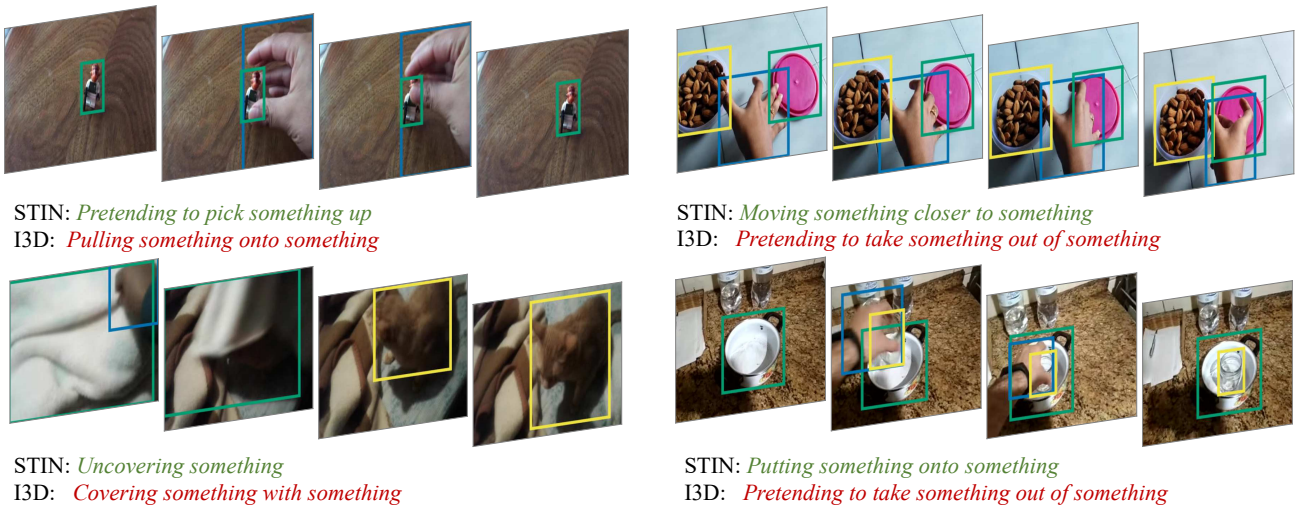


Figure 4. Predictions of STIN and I3D models. Correct predictions are in green, incorrect in red. STIN can keep tracking the relations between subjects and objects as they change over time in complicated actions.

as few-shot compositional recognition. We set k to 5 or 10 in our experiments. The statistics are shown in Table 1.

Bounding-box annotations. We annotated 180,049 videos of the Something-Something V2 dataset. For each video, we provide a bounding box of the hand (hands) and objects involved in the action. In total, 8,183,381 frames with 16,963,135 bounding boxes are annotated, with an average of 2.41 annotations per frame and 94.21 per video. Other large-scale video datasets use bounding box annotation, in applications involving human-object interaction [10], action recognition [21], and tracking [49].

5. Experiments

We perform experiments on the two proposed tasks: compositional action recognition and few-shot compositional action recognition.

5.1. Implementation Details

Detector. We choose Faster R-CNN [50, 71] with Feature Pyramid Network (FPN) [42] and ResNet-101 [24] backbone. The model is first pre-trained with the COCO [43] dataset, then finetuned with our object box annotations on the Something-Something dataset. During finetuning, only two categories are registered for the detector: *hand* and *object* involved in action. The object detector is trained with the same split as the action recognition model. We set the number of objects in our model as 4. If fewer objects are presented, we fill a zero vector to represent the object.

Tracker. Once we have the object detection results, we apply multi-object tracking to find correspondence between the objects in different frames. The multi-object tracker is implemented based on minimalism to keep the system as simple as possible. Specifically, we use the Kalman Filter [34] and Kuhn-Munkres (KM) algorithm [39] for tracking objects as [4]. At each time step, the Kalman Filter predicts plausible whereabouts of instances in the current frame based on previous tracks, then the predictions are matched with single-frame detections by the KM algorithm.

5.2. Setup

Training details. The MLP in our model contains 2 layers. We set the dimension of MLP outputs $d = 512$. We train all our models for 50 epochs with learning rate 0.01 using SGD with 0.0001 weight decay and 0.9 momentum, the learning rate is decayed by the factor of 10 at epochs 35 and 45.

Methods and baselines. The experiments aim to explore the effectiveness of different components in our Spatial-Temporal Interaction Networks for compositional action recognition, we compare the following models:

- **STIN**: Spatial-Temporal Interaction Network with bounding box coordinates as input. Average pooling is used as aggregation operator h.
- **STIN + OIE**: STIN model not only takes box coordinates but also Object Identity Embeddings (OIE).
- **STIN + OIE + NL**: Use non-local operators for aggregation operator h in STIN + OIE.
- **I3D**: A 3D ConvNet model with ResNet-50 backbone as in [68], with state-of-the-art performance.
- **STRG**: Space-Time Region Graph (STRG) model introduced in [68] with only similarity graph.
- **I3D + STIN + OIE + NL**: Combining the appearance feature from the I3D model and the feature from the STIN + OIE + NL model by joint learning.
- **I3D, STIN + OIE + NL**: A simple ensemble model combining the separately trained I3D model and the trained STIN + OIE + NL model.
- **STRG, STIN + OIE + NL**: An ensemble model combining the STRG model and the STIN + OIE + NL model, both trained separately.

Our experiments with STIN use either ground-truth boxes or the boxes detected by the object detector. The presented score is from a single clip in each video, which is a center cropped in time.

Visualization. Figure 4 visualizes examples of how our STIN model and I3D model performs. Our STIN model can keep tracking how the hand moves to understand the action whereas I3D is confused when the activity resembles other action class.

5.3. Original Something-Something Split

We first perform our experiments on the original Something-something V2 split. We test our I3D baseline model and the STIN model with ground-truth object bounding boxes for action recognition. As shown in Table 2, our I3D baseline is much better than the recently proposed TRN [77] model. The result of our STIN + OIE model with ground-truth annotations is reported in Table 2. We can see that with only coordinates inputs, our performance is comparable with TRN [77]. After combining with the I3D baseline model, we can improve the baseline model by 5%. This indicates the potential of our model and bounding box annotations even for the standard action recognition task.

5.4. Compositional Action Recognition

We further evaluate our model on the compositional action recognition task. We first experiment with using the ground-truth object bounding boxes for the STIN model, as reported in Table 3a. To illustrate the difficulty of our compositional task, we also report the results on a "shuffled" split of the videos: We use the same candidate videos but shuffle them randomly and form a new training and validation set. Note that the number of training videos is the same as the compositional split. The performance of the I3D baseline sharply drops from the shuffled setting to compositional setting by almost 15% in terms of top-1 accuracy. On the shuffled split, although our STIN model trails I3D, it performs better than I3D in the compositional split. By applying the Object Identity Embedding (OIE), we can improve the STIN model by 4.3%. This attests to the importance of explicit reasoning about the interactions between the agent and the objects. We can further combine our model with the I3D baseline: the joining of two models yields 7.8% improvement over the baseline and the ensemble model significantly improves over the appearance only model (I3D) by 11.3%.

Following, we build our model on object bounding boxes obtained via object detection and tracking and show its results in Table 3b. We observe that OIE still boosts the STIN model by 3.1%. By combining I3D with our model, we observe 1.4% improvement over I3D We also see that by replacing the base network with STRG we obtain some improvement in performance over I3D. After combining the STRG model with our model (STRG, STIN + OIE + NL), we can still achieve a large relative improvement (3.9% better than STRG). This shows that our method is complementary to the existing graph model.

model	top-1	top-5
TRN [77]	48.8	77.6
TRN Dual Attention [72]	51.6	80.3
TSM [41]	61.7	87.4
STIN + OIE	48.4	78.7
I3D	55.5	81.4
I3D + STIN + OIE	60.2	84.4

Table 2. Results on the **original Something-something V2** dataset. Ground-truth annotations are applied with STIN.

model	split	top-1	top-5
STIN	Shuffled	54.0	79.6
STIN	Compositional	47.1	75.2
STIN + OIE	Compositional	51.3	79.3
STIN + OIE + NL	Compositional	51.4	79.3
I3D	Shuffled	61.7	83.5
I3D	Compositional	46.8	72.2
I3D + STIN + OIE + NL	Compositional	54.6	79.4
I3D, STIN + OIE + NL	Compositional	58.1	83.2

(a) Compositional action recognition with **ground-truths**.

model	split	top-1	top-5
STIN	Compositional	34.1	58.8
STIN + OIE	Compositional	36.7	62.2
STIN + OIE + NL	Compositional	37.2	62.4
I3D	Compositional	46.8	72.2
STRG	Compositional	52.3	78.3
I3D + STIN + OIE + NL	Compositional	48.2	72.6
I3D, STIN + OIE + NL	Compositional	51.5	77.1
STRG, STIN + OIE + NL	Compositional	56.2	81.3

(b) Compositional action recognition with **detections**.

Table 3. **Compositional action recognition** over 174 categories.

5.5. Few-shot Compositional Action Recognition

For the few-shot compositional action recognition task, we have 88 *base* categories and 86 *novel* categories as described in Section 4. We first train our model with the videos from the base categories, then finetune on few-shot samples from the novel categories. We evaluate the model on the novel categories with more than 50k videos to benchmark the generalizability of our model. For finetuning, instead of following the n-way, k-shot setting in few-shot learning [16], we directly fine-tune our model with all the novel categories. For example, if we perform 5-shot training, then the number of training examples is $86\times5=430$. During the fine-tuning stage, we randomly initialize the last classification layer and train this layer while fixing all other layers. We train the network for 50 epochs with a fixed learning rate of 0.01. We perform both 5-shot and 10-shot learning in our experiment.

We report our results with ground-truth object boxes in Table 4a. We can see that our full model (STIN+OIE+NL) outperforms the I3D model by almost 6% in both 5-shot and 10-shot learning setting, even though our approach trails I3D on the validation set in *base* categories. This indicates

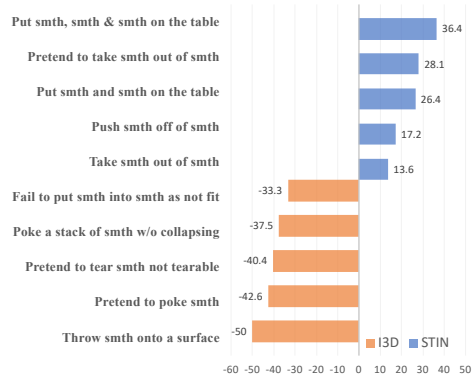


Figure 5. Top categories on which STIN surpasses or trails I3D, the numbers represent the difference in accuracy of the both models.

that the I3D representation can easily overfit to object appearance while our model generalizes much better. We also observe the OIE and non-local block individually and cooperatively boost the few-shot performance. When combining with I3D with the model ensemble, we achieve 12.2% improvement on 5-shot and 13.9% on the 10-shot setting. The results with object detection boxes are shown in Table 4b. Although the best model STIN+OIE+NL trails I3D on *base* evaluation by a notably large margin, the performance in the few-shot setting is much closer. When combining our model with the I3D model, joint learning yields 1.9% improvement and model ensemble yields 5.5% improvement in the 5-shot setting. We observe similar improvement in the 10-shot setting (5.9%). By replacing the I3D base network with STRG, our method (STRG, STIN + OIE + NL) still gives large improvement over STRG (4.3% in 5-shot setting and 4.7% in 10-shot).

5.6. Ablations

One-object training. We push the compositional setting to an extreme, where we only select the videos where actions are interacting with the object category "box" for training (6,560 videos in 166 action categories). The rest of the videos are the validation set (170K videos). The objective of this experiment is to examine the generalizability of our STIN model, even when the training set is strongly biased toward one type of object.

The results are summarized in Table 5. Our model with ground-truth boxes almost doubles the I3D performance. Our model with detection boxes is also 5.6% better than I3D. This attests to the advantage of our model in terms of generalizability across different object appearances.

Category analysis. We compare the performance difference between our STIN and the I3D model for individual action categories. We visualize the five action categories that STIN surpasses or trails by the largest margin compared to the I3D model in Figure 5. *A priori*, actions that are closely associated with the transformation of the object's geometric relations should be better represented by

model	*base* top-1	*base* top-5	*few-shot* 5-shot	*few-shot* 10-shot
STIN	65.7	89.1	24.5	30.3
STIN + OIE	69.5	91.4	25.8	32.9
STIN + OIE + NL	70.2	91.4	27.7	33.5
I3D	73.6	92.2	21.8	26.7
I3D + STIN + OIE + NL	80.6	95.2	28.1	33.6
I3D, STIN + OIE + NL	81.1	96.0	34.0	40.6

(a) Few-shot compositional setting with **ground-truths**.

model	*base* top-1	*base* top-5	*few-shot* 5-shot	*few-shot* 10-shot
STIN	54.0	78.9	14.2	19.0
STIN + OIE	58.2	82.6	16.3	20.8
STIN + OIE + NL	58.2	82.6	17.7	20.7
I3D	73.6	92.2	21.8	26.7
STRG	75.4	92.7	24.8	29.9
I3D + STIN + OIE + NL	76.8	93.3	23.7	27.0
I3D, STIN + OIE + NL	76.1	92.7	27.3	32.6
STRG, STIN + OIE + NL	78.1	94.5	29.1	34.6

(b) Few-shot compositional setting with **detections**.

Table 4. **Few-shot compositional action recognition** on *base* categories and *few-shot novel* categories. We show results with (a) ground-truth bounding boxes and (b) object detection boxes.

model	split	top-1	top-5
STIN + OIE (GT)	Compositional	28.5	54.1
STIN + OIE (Detector)	Compositional	20.3	40.4
I3D	Compositional	14.7	34.0

Table 5. **One-class** compositional action recognition. The model is trained on videos with only one object class: "box".

the STIN model than I3D. We can see that the actions in which STIN outperforms I3D by the largest margin are the ones that directly describe the movements of objects, such as *"put something"* and *"take something"*. On the other hand, STIN fails when actions are associated more with the changes in terms of the intrinsic property of an object, such as *"poking"* and *"tearing"*.

6. Conclusion

Motivated by the appearance bias in current activity recognition models, we propose a new model for action recognition based on sparse semantically grounded subject-object graph representations. We validate our approach on novel compositional and few shot settings in the Something-Else dataset; our model is trained with new constituent object grounding annotations Our STIN approach models the interaction dynamics of objects composed in an action and outperforms all baselines.

Acknowledgement: Prof. Darrell's group was supported in part by DoD, NSF, BAIR, and BDD. We would like to thank Fisher Yu and Haofeng Chen for helping set up the annotation pipeline, and Anna Rohrbach and Ronghang Hu for helpful discussions.

References

[1] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *CVPR*, pages 39–48, 2016. 3

[2] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. In *NIPS*, 2016. 3

[3] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. 2

[4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*, pages 3464–3468. IEEE, 2016. 6

[5] Yunlong Bian, Chuang Gan, Xiao Liu, Fu Li, Xiang Long, Yandong Li, Heng Qi, Jie Zhou, Shilei Wen, and Yuanqing Lin. Revisiting the effectiveness of off-the-shelf temporal modeling approaches for large-scale video classification. *arXiv:1708.03805*, 2017. 2

[6] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. *arXiv preprint arXiv:1906.11415*, 2019. 3

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2

[8] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *ICLR*, 2019. 3

[9] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *CVPR*, pages 433–442, 2019. 2

[10] 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 *ECCV*, pages 720–736, 2018. 6

[11] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 2

[12] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 2

[13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *arXiv preprint arXiv:1812.03982*, 2018. 2

[14] Christoph Feichtenhofer, Axel Pinz, and Richard Wildes. Spatiotemporal residual networks for video action recognition. In *NIPS*, 2016. 2

[15] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2009. 3

[16] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135. JMLR. org, 2017. 3, 7

[17] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*, 2017. 3

[18] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *CVPR*, pages 244–253, 2019. 3

[19] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object intaractions. *CVPR*, 2018. 3

[20] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 2, 5

[21] 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 *CVPR*, pages 6047–6056, 2018. 6

[22] Michelle Guo, Edward Chou, De-An Huang, Shuran Song, Serena Yeung, and Li Fei-Fei. Neural graph matching networks for fewshot 3d action recognition. In *ECCV*, pages 653–669, 2018. 3

[23] Abhinav Gupta, Aniruddha Kembhavi, and Larry S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *TPAMI*, 2009. 3

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[25] Roei Herzig, Elad Levi, Huijuan Xu, Hang Gao, Eli Brosh, Xiaolong Wang, Amir Globerson, and Trevor Darrell. Spatio-temporal action graph networks. In *ICCVW*, Oct 2019. 2, 3

[26] Donald D Hoffman and Whitman A Richards. Parts of recognition. *Cognition*, 18(1-3):65–96, 1984. 3

[27] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *CVPR*, 2018. 3

[28] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *CVPR*, pages 804–813, 2017. 3

[29] Nazlı İkizler and David A Forsyth. Searching for complex human activities with no visual examples. *IJCV*, 80(3):337–357, 2008. 3

[30] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *CVPR*, pages 5308–5317, 2016. 3

[31] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *TPAMI*, 2013. 2

[32] 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. In *ICCV*, pages 2989–2998, 2017. 3

[33] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *CVPR*, pages 3668–3678, 2015. 3

[34] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960. 6

[35] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732, 2014. 2

[36] Keizo Kato, Yin Li, and Abhinav Gupta. Compositional learning for human object interaction. In *ECCV*, pages 234–251, 2018. 3

[37] 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. 2

[38] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 2

[39] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 6

[40] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *CVPR*, pages 10657–10665, 2019. 3

[41] Ji Lin, Chuang Gan, and Song Han. Temporal shift module for efficient video understanding. *arXiv preprint arXiv:1811.08383*, 2018. 7

[42] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 6

[43] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 6

[44] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *ICCVW*, Oct 2019. 2

[45] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *CVPR*, pages 1792–1801, 2017. 3

[46] Ben Packer, Kate Saenko, and Daphne Koller. A combined pose, object, and feature model for action understanding. In *CVPR*, pages 1378–1385. IEEE, 2012. 3

[47] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. In *CVPR*, 2016. 2

[48] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2016. 3

[49] Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *CVPR*, pages 5296–5305, 2017. 6

[50] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 6

[51] Bryan C Russell, William T Freeman, Alexei A Efros, Josef Sivic, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006. 3

[52] Kate Saenko, Ben Packer, C Chen, S Bandla, Y Lee, Yangqing Jia, J Niebles, D Koller, L Fei-Fei, K Grauman, et al. Mid-level features improve recognition of interactive activities. *UCB EECS TR*, 2012. 3

[53] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *ICML*, pages 1842–1850, 2016. 3

[54] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *NIPS*, pages 4967–4976, 2017. 2

[55] Adam Santoro, David Raposo, David GT Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *NIPS*, 2017. 3

[56] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 2

[57] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 2

[58] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NIPS*, pages 4077–4087, 2017. 3

[59] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[60] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-centric relation network. In *ECCV*, pages 318–334, 2018. 3

[61] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 2

[62] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 2

[63] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 2

[64] Shubham Tulsiani, Hao Su, Leonidas J. Guibas, Alexei A. Efros, and Jitendra Malik. Learning shape abstractions by assembling volumetric primitives. In *CVPR*, 2017. 3

[65] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NIPS*, pages 3630–3638, 2016. 3

[66] 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 *ECCV*, 2016. 2

[67] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 2, 3, 4

[68] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, pages 399–417, 2018. 2, 3, 4, 6

[69] Nicholas Watters, Daniel Zoran, Theophane Weber, Peter Battaglia, Razvan Pascanu, and Andrea Tacchetti. Visual interaction networks: Learning a physics simulator from video. In *NIPS*, pages 4539–4547, 2017. 3

[70] 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. 3

[71] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 6

[72] Tete Xiao, Quanfu Fan, Dan Gutfreund, Mathew Monfort, Aude Oliva, and Bolei Zhou. Reasoning about human-object interactions through dual attention networks. In *CVPR*, pages 3919–3928, 2019. 7

[73] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 2

[74] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning for video understanding. In *arXiv:1712.04851*, 2017. 2

[75] Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 3

[76] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *CVPR*, 2015. 2

[77] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, pages 803–818, 2018. 2, 7

[78] Song-Chun Zhu, David Mumford, et al. A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision*, 2(4):259–362, 2007. 3