Video Relationship Reasoning using Gated Spatio-Temporal Energy Graph

Yao-Hung Hubert Tsai[†], Santosh Divvala[‡], Louis-Philippe Morency[†], Ruslan Salakhutdinov[†], Ali Farhadi[‡*]
[†]Carnegie Mellon University, [‡]Allen Institute for AI, *University of Washington
https://github.com/yaohungt/GSTEG_CVPR_2019

Abstract

Visual relationship reasoning is a crucial yet challenging task for understanding rich interactions across visual concepts. For example, a relationship {man, open, door} involves a complex relation {open} between concrete entities {man, door}. While much of the existing work has studied this problem in the context of still images, understanding visual relationships in videos has received limited attention. Due to their temporal nature, videos enable us to model and reason about a more comprehensive set of visual relationships, such as those requiring multiple (temporal) observations (e.g., {man, lift up, box} vs. {man, put down, box}), as well as relationships that are often correlated through time (e.g., {woman, pay, money} followed by {woman, buy, coffee}). In this paper, we construct a Conditional Random Field on a fully-connected spatio-temporal graph that exploits the statistical dependency between relational entities spatially and temporally. We introduce a novel gated energy function parametrization that learns adaptive relations conditioned on visual observations. Our model optimization is computationally efficient, and its space computation complexity is significantly amortized through our proposed parameterization. Experimental results on benchmark video datasets (ImageNet Video and Charades) demonstrate state-of-the-art performance across three standard relationship reasoning tasks: Detection, Tagging, and Recognition.

1. Introduction

Relationship reasoning is a challenging task that not only involves detecting low-level entities (subjects, objects, etc.) but also recognizing the high-level interaction between them (actions, sizes, parts, etc.). Successfully reasoning about relationships not only enables us to build richer question-answering models (e.g., *Which objects are larger than a car?*), but also helps in improving image retrieval [20](e.g., images with *elephants drawing a cart*), scene graph parsing [41] (e.g., *woman has helmet*), captioning [42], and many other visual reasoning tasks.

Most contemporary research in visual relationship reasoning has been focused in the domain of static images. While this has resulted in several exciting and attractive reasoning modules [26, 20, 42, 18, 40, 45, 3, 17], it lacks the ability from reasoning about complex relations that are inherently temporal and/or correlated in nature. For example, in Fig. 1 it is ambiguous to infer from a static image whether the monkey is creeping down or up the car. Also, it is difficult to model relations that are often correlated through time, such as *man enters room* and *man open door*.

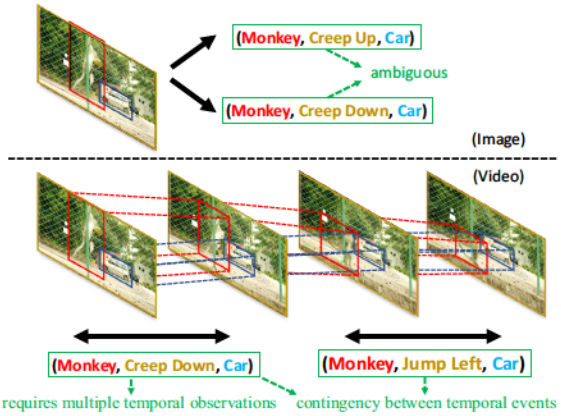


Figure 1. Visual relationship reasoning in images (top) vs. videos (bottom): Given a single image, it is ambiguous whether the *monkey* is creeping up or down the *car*. Using a video not only helps to unambiguously recognize a richer set of relations, but also model temporal correlations across them (e.g., *creep down* and *jump left*).

In this paper, we present a novel approach for reasoning about visual relationships in videos. Our proposed approach jointly models the spatial and temporal structure of relationships in videos by constructing a fully-connected spatio-temporal graph (see Fig. 2). We refer to our model as a Gated Spatio-Temporal Energy Graph. In our graph, each node represents an entity and the edges between them denote the statistical relations. Unlike much of the previous work [15, 43, 27, 4, 31] that assumed a predefined or globally-learned pairwise energy function, we introduce an observation-gated version that allows us to make the statistical dependency between entities adaptive (conditioned on the observation).

Our adaptive parameterization of energy function helps

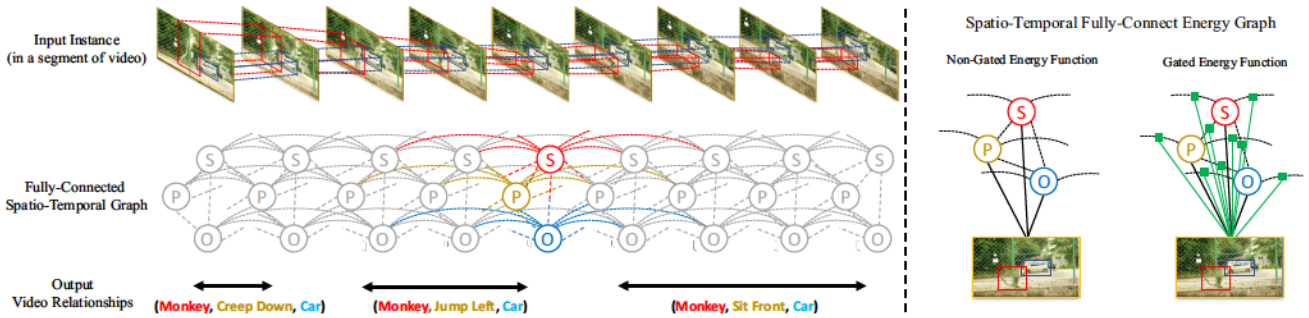


Figure 2. An overview of our Proposed Gated Spatio-Temporal Energy Graph. Given an input instance (a video clip), we predict the output relationships (e.g., {*monkey, creep down, car*}, etc.,) by reasoning over a fully-connected spatio-temporal graph with nodes S (Subject), P (Predicate) and O (Object). Unlike previous works that assumed a non-gated (i.e., predefined or globally-learned) pairwise energy function, we explore the use of gated energy functions (i.e., conditioned on the specific visual observation) . Best viewed zoomed in and in color.

us model the natural diversification of relationships in videos. For instance, the dependency between *man* and *cooking* should be different conditioned on the observation (i.e., whether the location is *kitchen* or *gym*). However, given the large state space of observations (in videos), directly maintaining observation-dependent statistical dependencies may be computationally intractable [22, 35]. Towards this end, we develop an amortized parameterization of our new gated pairwise energy function, which combines ideas from clique template [33, 34, 21], neural networks [8, 35], and tensor factorization [14] for achieving efficient inference and learning.

We evaluate our model on two benchmark datasets, ImageNet Video [24] and Charades [32]. Our method achieves state-of-the-art performance across three standard relationship reasoning tasks: detection, tagging, and recognition. We also study the utility of our model in the zero-shot setting and learning from semantic priors.

2. Related Work

Video Activity Recognition. The notion of activity in a video represents the interaction between objects [9, 12] or the interaction between an object and a scene [32]. While related to our task of relation reasoning, activity recognition does not require explicit prediction of all entities, such as subject, object, scene, and their relationships. The term *relation* used in activity recognition and relationship reasoning has different connotations. In the visual relationship reasoning literature, it refers to the correlation between different entities, such as object, verb, and scene, while in activity recognition, it refers to either correlation between activity predictions (i.e., single entity) or correlation between video segments. For example, [44] proposed Temporal Relation Network to reason the temporal 'relations' across frames at multiple time scales. [6] introduced a spatio-temporal aggregation on local convolutional features for better learning representations in the video. [38] proposed Non-Local Neural Networks to model pairwise relations for every pixel in the feature space from low-layers to higher-layers. The work was extended to [39] for constructing a Graph Convolutional Layer that further modeled relation between object-level features.

Visual Relationship Reasoning. Most recent works in relation reasoning have focused their analysis on static images [40, 45, 3, 17]. For example, [26] introduced the idea of visual phrases for compositing visual concepts of subject, predicate, and object. [20] decomposed the direct visual phrase detection task into individual detection on subject, predicate, and object leading to improved performance. [4] further applied conditional random fields on top of the individual predictions to leverage their statistical correlations. [18] proposed a deep variation-structured reinforcement learning framework and then formed a directed semantic action graph. The global interdependency in this graph facilitated predictions in local regions of the image. One of the key challenges of learning relationships in videos has been the lack of relevant annotated datasets. In this context, the recent work of [29] is inspiring as it contributes manually annotated relations for the ImageNet video dataset. Our work improves upon [29] on multiple fronts: (1) Instead of assuming no temporal contingency between relationships, we introduce a gated fully-connected spatio-temporal energy graph for modeling the inherently rich structure from videos; (2) We extend the study of relation triplet from subject/predicate/object to a more general setting, such as object/verb/scene [32]; (3) We consider a new task 'relation recognition' (apart from relation detection and tagging) which requires the model to make predictions in a fine-grained manner; (4) For various metrics and tasks, our model demonstrates improved performance.

Deep Conditional Random Fields. Conditional Random Fields (CRFs) have been popularly used to model the statistical dependencies among predictions in images [10, 43, 27, 25, 4] and videos [23, 31]. Several extensions have been recently introduced for fully-connected CRF graphs. For example, [43, 27, 31] attempted to express fully-connected CRFs as recurrent neural networks and made the whole net-

work end-to-end trainable, which has led to interesting applications in image segmentation [43, 27] and video activity recognition tasks [31]. In the characterization of CRFs, the unary energy function represents the inverse likelihood for assigning a label, while the binary energy function measures the cost of assigning multiple labels jointly. However, most of the existing parameterizations of binary energy functions [15, 43, 27, 4, 31] have limited or no connections to observed variables. Such parameterizations may not be optimal for video relationship reasoning due to the adaptive idiosyncrasy for statistical dependencies between entities. To address the issue, we instead propose an observation-gated pairwise energy function with efficient and amortized parameterization.

3. Proposed Approach

The task of video relationship reasoning not only requires modeling the entity predictions spatially and temporally, but also maintaining a changeable correlation structure between entities across videos with various contents. To this end, we propose a Gated Spatio-Temporal Fully-Connected Energy Graph for capturing the inherently rich video structure into account.

We start by defining our notations using Fig. 2 as a running example. The input instance X lies in a video segment and consists of K synchronous input streams $X = \{X^k\}_{k=1}^K$. In this example, input streams are {object trajectories, predicate trajectories, subject trajectories}, and thus $K = 3$, where trajectories refer to the consecutive frames or bounding boxes in the video segment. Each input stream contains observations for T time steps (i.e., $X^k = \{X_t^k\}_{t=1}^T$), where for example object trajectories represent object bounding boxes through time. For each input stream, our goal is to predict a sequence of entities (labels) $Y^k = \{Y_t^k\}_{t=1}^T$. In Fig. 2, the output sequence of predicate trajectories represent predicate labels through time. Hence we formulate the data-entities tuple as (X, Y) with $Y = \{Y_t^1, Y_t^2 \cdots, Y_t^K\}_{t=1}^T$ representing a set of sequence of entities.

The entity Y_t^k should spatially relate to entities $\{\{Y_t^1, Y_t^2 \cdots, Y_t^K\} \setminus \{Y_t^k\}\}$ and temporally relate to entities $\{\{Y_1^k, Y_2^k \cdots, Y_T^k\} \setminus \{Y_t^k\}\}$. For example, suppose that the visual relationships observed in a grocery store are {{mother, pay, money}, {infant, get, milk}, {infant, drink, milk}}; spatial correlation must exist between mother/pay/money and temporal correlation must exist between pay/get/drink. We also note that implicit correlation may also exist between Y_t^k and $Y_{t'}^{k'}$ for $t \neq t', k \neq k'$. Based on the structural dependencies between entities, we propose to construct a Spatio-Temporal Fully-Connected Energy Graph (see Sec. 3.1), where each node represents an entity and each edge denotes the statistical dependencies between the connected nodes. To further take account that the statistical dependency between "get" and "drink" may be different depending on different observations (i.e., location in grocery store v.s. home), we introduce an observation-gated parameterization for pairwise energy functions. In the new parameterization, we amortize the potentially large computational cost by using clique templates [33, 34, 21], neural network approximation [22, 35], and tensor factorization [14] (see Sec. 3.2).

3.1. Spatio-Temporal Fully-Connected Graph

By treating the predictions of entities as random variables, the construction of the graph can be realized by forming a Markov Random Field (MRF) conditioned on a global observation, which is the input instance (i.e., X). Then, the tuple (X, Y) can be modeled as a Conditional Random Field (CRF) parametrized by a Gibbs distribution of the form: $P\Big(Y = y|X\Big) = \frac{1}{Z(X)}\exp\Big(-\mathbf{E}(y|X)\Big)$, where $Z(X)$ is the partition function and $\mathbf{E}(y|X)$ is the energy of assigning labels $Y = y = \{y_t^1, y_t^2, \cdots, y_t^K\}_{t=1}^T$ conditioned on X. Assuming only pairwise cliques in the graph (i.e., $P(y|X) := P_{\psi,\varphi}(y|X), \mathbf{E}(y|X) := \mathbf{E}_{\psi,\varphi}(y|X)\big)$, the energy can be expressed as:

$$\mathbf{E}_{\psi,\varphi}(y|X) = \sum_{t,k} \psi_{t,k}(y_t^k|X) + \sum_{\{t,k\}\neq\{t',k'\}} \varphi_{t,k,t',k'}(y_t^k, y_{t'}^{k'}|X), \tag{1}$$

where $\psi_{t,k}$ and $\varphi_{t,k,t',k'}$ are the unary and pairwise energy, respectively. In Eq. (1), the unary energy, which is defined on each node in the graph, captures inverse likelihood for assigning $Y_t^k = y_t^k$ conditioned on the observation X. Typically, this term can be derived from an arbitrary classifier or regressor, such as a deep neural network [16]. On the other hand, the pairwise energy models interactions of label assignments across nodes $Y_t^k = y_t^k, Y_{t'}^{k'} = y_{t'}^{k'}$ conditioned on the observation X. Therefore, the pairwise term determines the statistical dependencies between entities spatially and temporally. However, the parameterization in most previous works on fully-connected CRF [43, 27, 31, 4] assumes that the pairwise energy function is non-adaptive to the current observation, which may not be ideal to model changeable dependencies between entities across videos. In the following Sec.3.2, we propose an observation-gated parametrization for pairwise energy function to address the issue.

3.2. Gated Pairwise Energy Function

Much of existing work uses a simplified parameterization of pairwise energy function and typically considers only the *smoothness* of the joint label assignment. For instance, in Asynchronous Temporal Field [31], $\varphi.(y_t^k, y_{t'}^{k'}|X)$ is defined as $\mu(y_t^k, y_{t'}^{k'})K(t, t')$, where μ represents the label compatibility matrix and $K(t, t')$ is an affinity kernel measurement which represents the dis-

counting factor between t and t'. Similarly, in the image segmentation domain [43, 27], $\varphi_\cdot(s_i, s_j|I)$ is defined as $\mu(s_i, s_j)K(I_i, I_j)$, where $s_{\{i,j\}}$ is the segment label and $I_{\{i,j\}}$ is the input feature for location $\{i, j\}$ in image I. In these models, the pairwise energy comprises an observation-independent label compatibility matrix followed by a spatio or temporal discounting factor. We argue that the parametrization of pairwise energy function should be more expressive. To this end, we define the pairwise energy as:

$$\varphi_{t,k,t',k'}(y_t^k, y_{t'}^{k'}|X) := \langle f^\varphi \rangle_{X,t,t',k,k',y_t^k,y_{t'}^{k'}}, \quad (2)$$

where f^φ can be seen as a discrete lookup table that takes the input X of size $|X|$ and outputs a large transition matrix of size $(T^2K^2 - 1) \times |Y_t^k| \times |Y_{t'}^{k'}|$, and where $\langle\cdot\rangle_z$ represents its z_{th} item. Directly maintaining this lookup table is computationally intractable due to the large state space of X. Considering a simple case that X is a pairwise-valued 32×32 image, we have $|X| = 2^{32\times32}$ possible states. The state space complexity aggravates when X becomes an RGB-valued video. Thanks to the recent advances in graphical models [33, 34, 21], deep neural networks [22, 35], and tensor factorization [14], our workaround is to parametrize and approximate f^φ as f_θ^φ with learnable parameters θ as follows:

$$\begin{aligned}\langle f^\varphi \rangle_{X,t,t',k,k',y_t^k,y_{t'}^{k'}} &\approx f_\theta^\varphi(X_t^k, t, t', k, k', y_t^k, y_{t'}^{k'}) \\ &= \begin{cases} \left\langle g_\theta^{kk'}(X_t^k) \otimes h_\theta^{kk'}(X_t^k) \right\rangle_{y_t^k, y_{t'}^{k'}} & t = t' \\ K_\sigma\left(t, t'\right) \left\langle r_\theta^{kk'}(X_t^k) \otimes s_\theta^{kk'}(X_t^k) \right\rangle_{y_t^k, y_{t'}^{k'}} & t \neq t' \end{cases}, \end{aligned} \quad (3)$$

where $g_\theta^{kk'}(\cdot), r_\theta^{kk'}(\cdot) \in \mathbb{R}^{|Y_t^k|\times r}$ and $h_\theta^{kk'}(\cdot), s_\theta^{kk'}(\cdot) \in \mathbb{R}^{|Y_{t'}^{k'}|\times r}$ represent the r-rank projection from X_t^k, which is modeled by a deep neural network. $A \otimes B = AB^\top$ denotes the function on matrix A and B, and results in a transition matrix of size $|Y_t^k| \times |Y_{t'}^{k'}|$. $K_\sigma\left(t, t'\right)$ is the Gaussian kernel with bandwidth σ representing discounting factor for different time steps.

The intuition behind our parametrization is as follows: First, we note that clique templates [33, 34, 21] are adopted spatially and temporally, which leads to scalable learning and inference. Second, the idea of using neural networks for approximating the lookup tables ensures both parameter efficiency and generalization [8, 35]. The lookup table maintains the state transitions of $\mathcal{X} \to \mathcal{Y}^k \times \mathcal{Y}^{k'}$ where calligraphy font denotes the corresponding state space. Finally, we choose $r << \min_k\{|Y_t^k|\}$ so that a low-rank decomposition is performed on the transition matrix from Y_t^k to $Y_{t'}^{k'}$. The low-rank decomposition allows us to substantially reduce the number of learnable parameters. To summarize, our design for f_θ^φ amortize the large space complexity for f^φ and is gated by observation.

3.3. Inference, Message Passing, and Learning

Minimizing the CRF energy in Eq. (1) returns the most probable label assignment problem of $Y = \{y_t^1, y_t^2, \cdots, y_t^K\}_{t=1}^T$ given the observation X. However, the exact inference in a fully connected CRF is often computationally intractable even with variables enumeration or elimination [13]. In this work, we adopt the commonly used mean-field algorithm [13] as approximate inference, which finds the approximate posterior distribution $Q(Y)$ such that $Q(\cdot)$ is closest to $P_{\psi,\varphi}(Y|X)$ in terms of $\mathcal{KL}(Q//P_{\psi,\varphi})$ within the class of distributions representable as a product of independent marginals $Q(Y) = \prod_{t,k} Q(Y_t^k)$. Following [13], inference can now be realized as the naive mean-field updates with the coordinate descent optimization, and it can be expressed in terms of fixed-point message passing equations:

$$Q(y_t^k) \propto \Psi_{t,k}\left(y_t^k|X\right) \prod_{\{t',k'\}\neq\{t,k\}} m_{t',k',t,k}(y_t^k|X) \quad (4)$$

with $\Psi_{t,k} = \exp\Big(-\psi_{t,k}\Big)$ representing the unary potential and $m_\cdot(\cdot)$ denoting the message having form[1] of

$$m_\cdot(\cdot) = \exp\Big(-\sum_{y_{t'}^{k'}} \varphi_{t,k,t',k'}(y_t^k, y_{t'}^{k'}|X)Q(y_{t'}^{k'})\Big). \quad (5)$$

To parametrize the unary energy function, we use a similar formulation:

$$\begin{aligned}\psi_{t,k}(y_t^k|X) &:= \left\langle f^\psi \right\rangle_{X,t,k,y_t^k} \\ &\approx f_\theta^\psi(X_t^k, t, k, y_t^k) = \left\langle w_\theta^k(X_t^k)\right\rangle_{y_t^k}, \end{aligned} \quad (6)$$

where $w_\theta^k \in \mathbb{R}^{|Y_t^k|}$ represents the projection from X_t^k to logits of size $|Y_t^k|$, modeled by a deep neural network.

Lastly, we cast the learning problem as minimizing conditional cross-entropy between the proposed distribution and the true one, where θ denotes the parameters we need in our model: $\theta^* = \arg\min_\theta \mathbb{E}_{X,Y}[-\log Q(Y)]$.

4. Experimental Results & Analysis

In this section, we report our quantitative and qualitative analyses for validating the benefit of our proposed method. Our experiments are designed to compare different baselines and ablations for detecting and tagging relationships given a video as well as recognizing relationships in a fine-grained manner.

[1] In Supplementary, we make connection from our gated amortized parametrization for pairwise energy function in message form with Self-Attention [36] in machine translation and Non-Local Means [1] in Image Denoising.

Method	Correponding Image-Relationship or Video-Activity Method	Relationship Detection relationship R@50	R@100	mAP	Relationship Tagging relationship P@1	P@5	P@10	Relationship Recognition subject Acc@1	predicate Acc@1	object Acc@1	relationship Acc@1
Standard Evaluation											
VidVRD* [29]	Visual Phrases [26]	5.58	6.68	6.94	41.00	29.60	21.85	80.28	16.55	80.40	12.93
UEG	VRD_V [20]	2.81	3.64	2.94	31.50	19.88	14.98	80.15	23.95	80.55	18.62
UEG†	VRD [20]	3.41	4.05	4.52	36.00	21.60	15.41	80.15	25.92	80.55	22.47
SEG	DRN [4]	4.34	5.32	4.16	35.00	27.10	20.90	85.15	25.85	84.26	20.97
STEG	AsyncTF [31]	4.18	4.98	4.71	40.00	24.45	17.66	89.91	25.92	89.33	22.54
GSTEG (Ours)	-	**7.05**	**8.67**	**9.52**	**51.50**	**39.50**	**28.23**	**90.60**	**28.78**	**89.79**	**25.01**
Zero-Shot Evaluation											
VidVRD* [29]	Visual Phrases [26]	0.93	1.16	**0.18**	0.0	0.82	0.82	74.54	2.78	74.07	1.62
UEG	VRD_V [20]	0.0	0.23	1.30×1e-5	0.0	0.27	0.82	74.31	5.09	74.77	3.24
UEG†	VRD [20]	0.23	0.23	5.36×1e-5	0.0	0.82	0.82	78.24	5.79	78.47	3.47
SEG	DRN [4]	0.23	0.46	6.70×1e-5	0.0	0.82	1.23	81.02	6.94	74.54	3.47
STEG	AsyncTF [31]	0.23	0.69	0.02	1.37	1.10	0.96	80.09	7.18	79.17	4.40
GSTEG (Ours)	-	**1.16**	**2.08**	0.15	**2.74**	**1.92**	**1.92**	**82.18**	**7.87**	**79.40**	**6.02**

Table 1. Evaluation for different methods on ImageNet Video dataset. * denotes the re-implementation of [29] after fixing the bugs in their released method code (by contacting authors). † denotes the implementation with additional triplet loss term for language priors [20].

Datasets. We perform our analysis on two datasets: ImageNet Video [24] and Charades [32]. (a) *ImageNet Video* [24] contains videos (from daily-life as well as in-the-wild) with manually labeled bounding boxes for objects. We utilize the annotations from [29], in which a subset of the videos having rich visual relationships were selected (1,000 videos in total with 800 for training & rest for evaluation, available at [28]). The visual relationship is defined on the triplet {*subject*, *predicate*, *object*}. For example, {*person*, *ride*, *bicycle*} or {*dog*, *larger*, *monkey*}, etc. It has 35 categories of *subject* and *object*, and 132 categories of *predicate* (see Suppl. for details) with trajectory denoting consecutive bounding boxes through time. A relation triplet is labeled on a pair of trajectories (one for subject and another for object). The entire video has multiple pairs of trajectories and these pairs may or may not overlap with each other spatially or temporally. (b) *Charades* [32] contains videos of human indoor activities (9,848 in total with 7,985 for training and the rest for evaluation, available at [30]) . The visual relationship is defined on the triplet {*verb*, *object*, *scene*} or {*object*, *verb*, *scene*}. For example, {*hold*, *blanket*, *bedroom*}, {*someone*, *cook*, *kitchen*}, etc. It has 33 categories of *verb*, 38 *objects* and 16 *scenes* (see Suppl. for details). Different from ImageNet Videos, as suggested by [32, 31, 38, 39], we treat the entire video as an input instance. Therefore, a video comprises multiple relation triplets, and each relation triplet is defined within a time segment. The relation triplets may or may not overlap temporally with each other.

Tasks. For the above two datasets, we consider the following three experimental tasks.

(i) Relationship Detection. For ImageNet Videos, we aim at predicting a set of visual relationships with estimated subject and object trajectories. Specifically, a predicted visual relationship is counted as correct if the predicted triplet is in the ground truth set and the estimated bounding boxes have high voluminal intersection over union (vIoU) with the ground truth (vIoU threshold of 0.5). Following [29], this task is termed *relationship detection*, which contains both relationship prediction and object localization. For Charades dataset, as suggested by [31][2], we aim at detecting the visual relationships in a video without object localization, i.e., relationship detection happens in the scale of the entire video. For evaluation, we follow [20, 29] and adopt mean average precision (mAP) and Recall@K (K equals 50 and 100) metrics, where mAP measures the average of the maximum precisions at different recall values and Recall@K measures the fraction of the positives detected in the top K detection results.

(ii) Relationship Tagging. For ImageNet Videos, the *relationship tagging* task [29] focuses on only relationship prediction. This is motivated by the fact that video object localization is still an open problem. Similarly, in Charades, relationship tagging focuses on only relationship prediction (where relationship tagging happens at the scale of entire video). Following [29], we use Precision@K (K equals 1, 5, and 10) to measure the accuracy of the tagging results.

(iii) Relationship Recognition. Different from performing relationship reasoning at the scale of entire video, we would also like to measure how well the model recognizes the relationship in a fine-grained manner. For example, given an object trajectory and a subject trajectory, can the model predict accurate relationships? For the ImageNet Video experiments: given an input instance (with object and subject trajectories in a time segment), we measure the recognition accuracy of subject, predicate, object, and the relationship, which we term it *relationship recognition*. As the Charades dataset does not consider object localization,

[2]The performance reported in [31] refers to the mean Avergage Precision (mAP) of 157 activities, while ours consider the detection of relation triplets. Although not being our focus, our method with the 157 activities output achieves 33.3 mAP on activity detection as compared to 18.3 mAP in [31] when using only RGB frames as input. See Suppl. for details.

Method	Correponding Image-Relationship or Video-Activity Method	Relationship Detection relationship			Relationship Tagging relationship			Relationship Recognition object	verb	scene	relationship
		R@50	R@100	mAP	P@1	P@5	P@10	Acc@1	Acc@1	Acc@1	Acc@1
VidVRD [29]	Visual Phrases [26]	13.62	18.36	3.12	3.97	4.62	4.26	28.70	63.64	34.91	7.83
UEG	VRD_V [20]	22.53	29.70	7.93	16.05	11.47	8.72	41.74	64.70	34.62	11.94
UEG†	VRD [20]	22.35	29.65	7.90	16.10	11.38	8.67	41.70	64.73	35.17	11.85
SEG	DRN [4]	23.68	31.56	8.77	18.04	12.50	9.37	42.84	64.36	35.28	12.60
STEG	AsyncTF [31]	23.79	31.65	8.84	18.46	12.57	9.37	42.87	64.53	35.71	12.76
GSTEG (Ours)	-	**24.95**	**33.37**	**9.86**	**19.16**	**12.93**	**9.55**	**43.53**	**64.82**	**40.11**	**14.73**

Table 2. Evaluation for different methods on Charades dataset. Our method outperforms all competing baselines across the three tasks.

we perform recognition on object, verb, scene, and the relationship within a time segment (where relation recognition happens at the scale of a time segment in the video). We use Accuracy@K (K equals 1) for emphasizing whether the model gives the correct recognition result on the top 1 relationship prediction.

Pre-Reasoning Modules For all our experiments and ablation studies, we use the following three (exactly same) pre-reasoning modules:

∘ *Video Chunking.* As suggested by [2], we treat the video as consecutive overlapping segments with each segment comprising continuous frames. For ImageNet Video, each segment contains 30 frames, and adjacent segments have 15 overlapping frames. Since the video is chunked, the object and subject trajectories are also decomposed into chunks. For Charades, each segment contains 10 frames, and adjacent segments have 6 overlapping frames.

∘ *Tracklet Proposal.* Tracklet proposal is required in the ImageNet Video dataset for object localization. For each chunk in the video, we generate proposals for the possible subject and object tracklets. We utilize Faster-RCNN [7] as object detector trained on the 35 objects (categories in the annotation) from MS-COCO [19] and ImageNet Detection [24] datasets. Next, the method described in [5] is used to relate frame-level into a chunk-level object proposals. Then, non-maximum suppression (NMS) with vIoU > 0.5 is performed to reduce the numbers of generated chunk-level proposals. During training, proposals that have vIoU > 0.5 with the ground truth trajectories are selected to be the training proposals. However, all the generated proposals are preserved for evaluation.

∘ *Feature Representation.* Following Sec. 3 notation, we express the input instance X into K synchronous streams of features. For the ImageNet Video, K equals 3 and the synchronous streams of features are $\{X_t^s, X_t^p, X_t^o\}_{t=1}^T$. s, p, o and T denote subject, predicate, object, and the number of chunks in the input instance, respectively. Note that each instance may have different numbers of chunks, i.e., different T, because of various duration of relationships. The output Y_t^s, Y_t^p , and Y_t^o follow categorical distribution. As in [29], in the t_{th} chunk of the input instance, we choose the subject and object features (i.e., X_t^s and X_t^o) to be the averaged features for the Faster-RCNN label probability distribution outputs. X_t^p, on the other hand, is chosen to be the concatenation of the following three features: the improved dense trajectory (iDT) feature [37] for subject tracklet, the iDT feature for object tracklet, and the relative spatio-temporal positions [29] between subject and object tracklets. See Suppl. for more details.

For Charades, the input instance X is expressed as $\{X_t^o, X_t^v, X_t^s\}_{t=1}^T$ with o, v, and s denoting object, verb, and scene , respectively. Since we are performing relationship reasoning directly in the entire video, we let Y_t^o, Y_t^v be a multinomial distribution while Y_t^s still remains to be a categorical distribution. The multinomial distribution suggests that each chunk may contain ≥ 0 number of objects or verbs. We set X_t^o, X_t^v, and , X_t^s to have identical features: the output feature layer from I3D network [2]. See Suppl. for more details.

Baselines The closest baseline to our proposed model is VidVRD [29]. Beyond comparisons to [29], we also perform a detailed ablation study of our method as well as relate to the image-based visual relationship reasoning methods (when applicable).

VidVRD. VidVRD [29] adopted a structured loss on the multiplication of three features (i.e., X_t^s, X_t^p, and X_t^o for ImageNet Video). The loss took softmax over all training triplets, which resembles the training objective in Visual Phrases [26] (designed for image-based visual relationship reasoning). Note that VidVRD fails to consider the temporal structure of relationship predictions.

GSTEG (Ours). We denote our proposed method as GSTEG (**G**ated **S**patio-**T**emporal **E**nergy **G**raph). For the ablation study, we choose the Energy Graph (EG) when considering different energy function designs as described below.

STEG. **S**patio-**T**emporal **E**nergy **G**raph (STEG) takes into account the spatial and temporal structure of video entities. However, it assumes fixed statistical dependencies between entities. Specifically, it is the non-gated version of our full model. STEG can be seen as a modified version of Asynchronous Temporal Fields (AsyncTF) [31] such that we have (1) AsyncTF's output to be a relationship prediction, and (2) a fully-connected spatial graph.

SEG. Compared to STEG, the **S**patio **E**nergy **G**raph (SEG) method does not consider the temporal structure of video entities. Specifically, SEG assumes a spatially-fully-connected graph and thus the relationship predictions are made temporally independently. The counterpart in image-

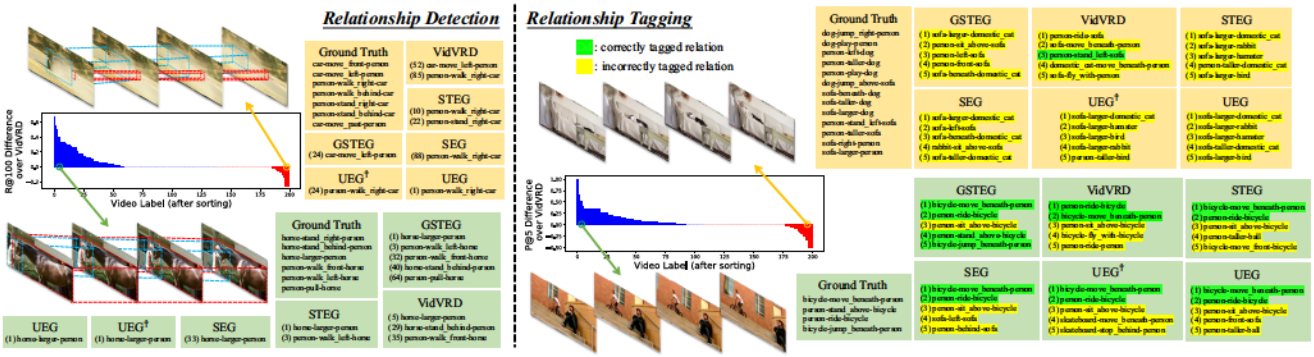


Figure 3. Examples from ImageNet Video dataset of Relationship Detection (Left) & Tagging (Right) using baselines, ablations, and our full model. The bar plots illustrate the R@100 (left) and P@5 (right) difference comparing our model to VidVRD [29]. To show the results on all the methods, green boxes refer to a video where our model performs better and orange boxes refer to a video where VidVRD performs better. For tagging (right), we use green to highlight the correctly tagged relation and yellow for incorrectly tagged relation. The numbers in bracket represent the order of detection or tagging. Best viewed in color.

based visual relationship reasoning methods is Deep Relational Networks (DRN) [4]. We can view SEG as casting DRN to (1) take the video-based input features and (2) consider continuous object bounding boxes through time instead of a bounding box in a single frame.

UEG and UEG†. The **U**nary **E**nergy **G**raph (UEG) considers the prediction of entities both spatially and temporally independently. The counterpart in image-based visual relationship reasoning methods is the Visual Relationship Detection (VRD) method of [20] without using language priors (denoted as VRD_V). Similar to the modification from DRN to SEG, the accommodation from VRD_V to UEG is having VRD_V take the video-based features and consider object trajectories. We also perform experiments that extend UEG with additional triplet loss defined with language priors [20], which we denote it as UEG†. The counterpart in image-based methods is the full VRD model of [20]. (Please see Suppl. for more details about parameterizations and training for all the methods and datasets).

4.1. Quantitative Analysis

ImageNet Video. Table. 1 shows our results and comparisons to the baselines. We first observe that, for every metric across the three tasks (detection, tagging, and recognition), our proposed method (GSTEG) outperforms all the competing methods. Comparing the numbers between UEG and UEG†, we find that language priors can help promote visual relation reasoning. We also observe performance improvement from UEG to SEG, which could be explained by the fact that SEG explicitly models the spatial statistical dependency in {subject, predicate, object} and leads to a better relation learning between different entities. However, comparing SEG to STEG, the performance drops in some metrics, indicating that modeling temporal statistical dependency using a fixed pairwise energy parameterization may not be ideal. For example, although STEG gives a much better relationship recognition results as compared to SEG, it becomes worse in R@50 for detection and P@5 for tagging. This indicates that observation-gated parametrization for pairwise energy is able to capture different structure for different videos. When comparing energy graph models, VidVRD is able to outperform all our ablation baselines (except for the full version) in relation detection and tagging. However, it suffers from relation recognition, which requires a fine-grained understanding of visual relation in the given object and subject tracklets.

Apart from the 'standard evaluation', we also considered the 'zero-shot' setting, where *zero-shot* refers to the evaluation on the relative complement of training triplets in evaluation triplets. More specifically, in the ImageNet Video dataset, the number of all possible relation triplets is $35 \times 132 \times 35 = 161,700$. While the training set contains $2,961$ relation triplets (i.e., 1.83% of $161,700$), the evaluation set has $1,011$ relation triplets (i.e., 0.63% of $161,700$). The number of zero-shot relation triplets is 258, which is 25.5% in the evaluation set. Zero-Shot evaluation is very challenging due to the fact that we need to infer the never-seen relationship in the training set. We observe that, for most cases, our proposed method reaches the best performance compared to various baselines. The exception is mAP, where VidVRD attains the best performance using a structural objective. However, the overall trend of zero-shot evaluation mirrors standard evaluation.

Charades. Our results and comparisons are shown in Table. 2. We find that our method outperforms all relevant baselines. We also note some interesting differences between the trend of results in Charades vs. ImageNet Video: First, comparing UEG to UEG†, we observe that language priors do not really help the visual relationship reasoning in Charades. We argue that it may because of the larger inter-class distinction in Charades' categories set. For example, dog/cat or horse/zebra or sit front/front/jump front share some similarity in the category set in ImageNet Video,

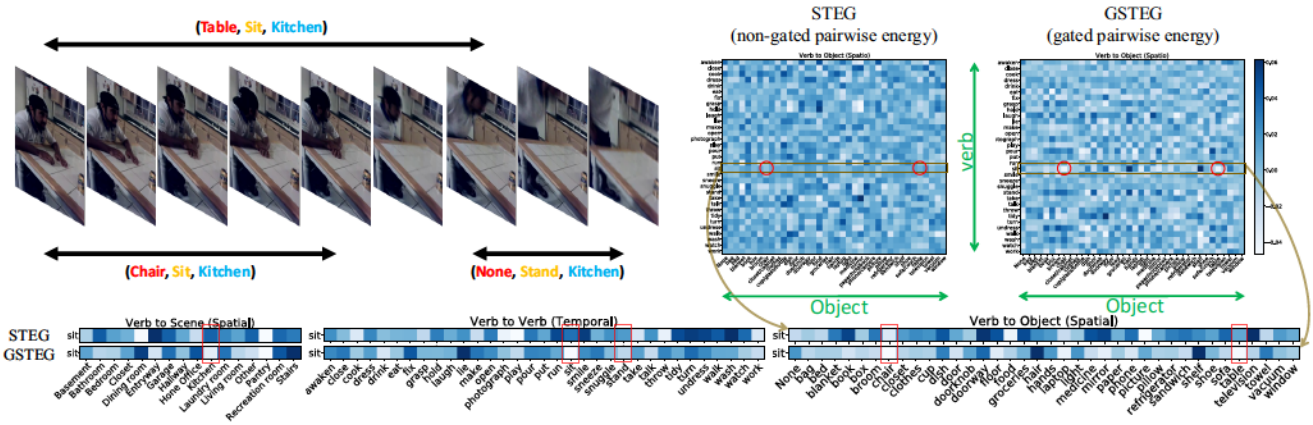


Figure 4. Analysis of non-gated and gated pairwise energies: Given an input video (top left) from Charades (that has {*object, verb, scene*} relationships), the matrices (top right) visualize the non-gated and gated pairwise energies between the verbs and objects (rows: 33 verbs, cols: 38 objects). Notice that for the verb *sit* (highlighted in red), the gated energy with objects *chair*, and *table* is lower compared to the corresponding non-gated pairwise energies, thereby helping towards improved relationship reasoning. A similar behavior is observed in case of verb to scene pairwise function (bottom left) as well as verb to verb pairwise function (bottom middle), which models the temporal correlations e.g., sit/sit or sit/stand. Best viewed in color and color in the matrix or vector is normalized in its own scale.

while the categories are less semantically similar in Charades. Second, STEG constantly outperforms SEG which indicates modeling a fixed temporal statistical dependency between entities may aid the visual relationship reasoning in Charades. We hypothesize that, as compared to the ImageNet Video dataset that has a diversified set of videos in the wild between animals or inorganic substances, Charades contains videos of human indoor activities where relations between entities are much easier to model by a fixed dependency. Finally, we observe that VidVRD performs substantially worse compared to all the other models, suggesting that the structural loss introduced by VidVRD may not generalize well to other datasets. In case of Charades, we do not perform zero-shot evaluation as the number of zero-shot relation triplets is low. (The number of all the possible relation triplets is $33 \times 38 \times 16 = 20,064$. The training set contains $2,285$ relation triplets (i.e., 11.39% of $20,064$) and the evaluation set contains $1,968$ relation triplets (i.e., 9.81% of $20,064$). The number of zero-shot relation triplets is 46, which is 2.34% in the evaluation set.)

In Supplementary, we also provide the results when leveraging language priors into our model and also provide the comparisons with Structural-RNN [11] and Graph Convolutional Network [39].

4.2. Qualitative Analysis

We next illustrate our qualitative results in Fig. 3 in the ImageNet Video dataset. For the relationship detection, in a scene with a person interacting with a horse, our model successfully detects 5 out of 6 relationships, while failing to detect horse-stand_right-person in the top 100 detected relationships. In another scene with a car interacting with a person, our model only detects 1 relationship out of 7 ground-truth relationships. We argue that the reason may be because of the sand occlusion and the small size of a person. For relationship tagging, in a scene with a person riding a bike over another person, our model successfully tags all four relationships in the top 5 tagged results. Nevertheless, the third tagged result person-sit_above-bicycle also looks visually plausible in this video. In another scene with a person playing with a dog on a sofa, our model fails to tag any correct relationships in the top 5 tagged results. Our model incorrectly identified dog as cat, representing the main reason why it failed.

Since pairwise energy in a graphical model represents the negative statistical dependency between entities, in Fig. 4, for a video in Charades dataset, we provide the illustration of pairwise energy when considering our gated and non-gated parameterization. Observe that the pairwise energies between the related entities are lower for the gated parameterization as compared to the non-gated one, suggesting that the gating mechanism is able to aid video relationship reasoning by improving statistical dependency between spatially or temporally correlated entities.

5. Conclusion

In this paper, we have presented a Gated Spatio-Temporal Energy Graph (GSTEG) model for the task of visual relationship reasoning in videos. In the graph, we consider a spatially and temporally fully-connected structure with an amortized observation-gated parameterization for the pairwise energy functions. The gated design enables the model to detect adaptive relations between entities conditioned on the current observation (i.e., current video). On two benchmark video datasets (ImageNet Video and Charades), our method achieves state-of-the-art performance across three relationship reasoning tasks (Detection, Tagging, and Recognition).

Acknoledgement

Work done when YHHT was in Allen Institute for AI. YHHT and RS were supported in part by the DARPA grants D17AP00001 and FA875018C0150, NSF IIS1763562, and Office of Naval Research N000141812861. LPM was supported by the National Science Foundation (Award # 1722822). SD and AF were supported by NSF IIS-165205, NSF IIS-1637479, NSF IIS-1703166, Sloan Fellowship, NVIDIA Artificial Intelligence Lab, and Allen Institute for artificial intelligence. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of National Science Foundation, and no official endorsement should be inferred. We would also like to acknowledge NVIDIAs GPU support.

References

[1] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 60–65. IEEE, 2005. 4

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4724–4733. IEEE, 2017. 6

[3] Zhen Cui, Chunyan Xu, Wenming Zheng, and Jian Yang. Context-dependent diffusion network for visual relationship detection. *arXiv preprint arXiv:1809.06213*, 2018. 1, 2

[4] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3298–3308. IEEE, 2017. 1, 2, 3, 5, 6

[5] Martin Danelljan, Gustav Häger, Fahad Khan, and Michael Felsberg. Accurate scale estimation for robust visual tracking. In *British Machine Vision Conference, Nottingham, September 1-5, 2014*. BMVA Press, 2014. 6

[6] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *CVPR*, volume 2, page 3, 2017. 2

[7] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 6

[8] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016. 2, 4

[9] 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 *The IEEE International Conference on Computer Vision (ICCV)*, volume 1, page 3, 2017. 2

[10] Xuming He, Richard S Zemel, and Miguel Á Carreira-Perpiñán. Multiscale conditional random fields for image labeling. In *Computer vision and pattern recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE computer society conference on*, volume 2, pages II–II. IEEE, 2004. 2

[11] 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. 8

[12] 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

[13] Daphne Koller, Nir Friedman, and Francis Bach. *Probabilistic graphical models: principles and techniques*. MIT press, 2009. 4

[14] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009. 2, 3, 4

[15] 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. 1, 3

[16] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015. 3

[17] Kongming Liang, Yuhong Guo, Hong Chang, and Xilin Chen. Visual relationship detection with deep structural ranking. 2018. 1, 2

[18] 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. 1, 2

[19] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[20] 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. 1, 2, 5, 6, 7

[21] Andrew McCallum, Karl Schultz, and Sameer Singh. Factorie: Probabilistic programming via imperatively defined factor graphs. In *Advances in Neural Information Processing Systems*, pages 1249–1257, 2009. 2, 3, 4

[22] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013. 2, 3, 4

[23] Ariadna Quattoni, Sybor Wang, Louis-Philippe Morency, Morency Collins, and Trevor Darrell. Hidden conditional random fields. *IEEE transactions on pattern analysis and machine intelligence*, 29(10), 2007. 2

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and

Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 2, 4, 6

[25] Fereshteh Sadeghi, Santosh K Kumar Divvala, and Ali Farhadi. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1456–1464, 2015. 2

[26] Mohammad Amin Sadeghi and Ali Farhadi. Recognition using visual phrases. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1745–1752. IEEE, 2011. 1, 2, 5, 6

[27] Alexander G Schwing and Raquel Urtasun. Fully connected deep structured networks. *arXiv preprint arXiv:1503.02351*, 2015. 1, 2, 3, 4

[28] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. *URL for ImageNet Video*. https://lms.comp.nus.edu.sg/research/VidVRD.html. 5

[29] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1300–1308. ACM, 2017. 2, 5, 6, 7

[30] Gunnar A Sigurdsson, Santosh Kumar Divvala, Ali Farhadi, and Abhinav Gupta. *URL for Charades*. http://ai2-website.s3.amazonaws.com/data/Charades.zip. 5

[31] Gunnar A Sigurdsson, Santosh Kumar Divvala, Ali Farhadi, and Abhinav Gupta. Asynchronous temporal fields for action recognition. In *CVPR*, volume 5, page 7, 2017. 1, 2, 3, 5, 6

[32] 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. 2, 4, 5

[33] Ben Taskar, Pieter Abbeel, and Daphne Koller. Discriminative probabilistic models for relational data. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 485–492. Morgan Kaufmann Publishers Inc., 2002. 2, 3, 4

[34] Graham W Taylor and Geoffrey E Hinton. Factored conditional restricted boltzmann machines for modeling motion style. In *Proceedings of the 26th annual international conference on machine learning*, pages 1025–1032. ACM, 2009. 2, 3, 4

[35] Yao-Hung Hubert Tsai, Han Zhao, Ruslan Salakhutdinov, and Nebojsa Jojic. Discovering order in unordered datasets: Generative markov networks. *arXiv preprint arXiv:1711.03167*, 2017. 2, 3, 4

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 4

[37] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013. 6

[38] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 5

[39] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. *arXiv preprint arXiv:1806.01810*, 2018. 2, 5, 8

[40] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Chen Change Loy. Zoom-net: Mining deep feature interactions for visual relationship recognition. *arXiv preprint arXiv:1807.04979*, 2018. 1, 2

[41] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018. 1

[42] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *CVPR*, volume 1, page 5, 2017. 1

[43] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537, 2015. 1, 2, 3, 4

[44] Bolei Zhou, Alex Andonian, and Antonio Torralba. Temporal relational reasoning in videos. *arXiv preprint arXiv:1711.08496*, 2017. 2

[45] Yaohui Zhu and Shuqiang Jiang. Deep structured learning for visual relationship detection. 2018. 1, 2

Supplementary for Video Relationship Reasoning using Gated Spatio-Temporal Energy Graph

1. Towards Leveraging Language Priors

The work of [5] has emphasized the role of language priors in alleviating the challenge of learning relationship models from limited training data. Motivated by their work, we also study the role of incorporating language priors in our framework.

In Table. 1 in the main text, comparing UEG to UEG†, we have seen that language priors aid in improving the relationship reasoning performance. Considering our example in Sec. 3.1 in main text, when the training instance is {mother, pay, money}, we may also want to infer that {father, pay, money} is a more likely relationship as opposed to {cat, pay, money} (as mother and father are semantically similar compared to mother and cat). Likewise, we can also infer {mother, pay, check} from the semantic similarity between *money* and *check*.

[5] adopted a triplet loss for pairing word embeddings of object, predicate, and subject. However, their method required sampling of all the possible relationships and was also restricted to the number of entities spatially (e.g, $K = 3$). Here, we present another way to make the parameterized pairwise energy also be gated by the prior knowledge in semantic space. We let the prior from semantic space be encoded as word embedding: $S = \{S^k\}_{k=1}^K$ in which $S^k \in \mathbb{R}^{|Y_t^k| \times d}$ denoting prior of labels with length d. We extend Eq. (3) in the main text as

$$\begin{aligned} &f_\theta^\varphi(S, X_t^k, t, t', k, k', y_t^k, y_{t'}^{k'}) \\ =&f_\theta^\varphi(X_t^k, t, t', k, k', y_t^k, y_{t'}^{k'}) + u_\theta(\left\langle S^k \right\rangle_{y_t^k}) \cdot v_\theta(\left\langle S^{k'} \right\rangle_{y_{t'}^{k'}}), \end{aligned} \tag{1}$$

where $u_\theta(\cdot) \in \mathbb{R}$ and $v_\theta(\cdot) \in \mathbb{R}$ maps the label prior to a score. Eq. (1) suggests that the label transition from Y_t^k to $Y_{t'}^{k'}$ can also attend to the affinity inferred from prior knowledge.

We performed a preliminary evaluation on the relation recognition task in the ImageNet Video dataset using 300-dim Glove features [6] as word embeddings. For subject, predicate, object, and relation triplet, Acc@1 metric improves from 90.60, 28.78, 89.79, and 25.01 to 90.97, 29.54, 90.57, and 26.48.

2. Connection to Self Attention and Non-Local Means

In our main text, the message form (eq. (5)) with our observation-gated parametrization (eq. (3) with $t = t'$) can be expressed as follows:

$$\begin{aligned} -\log m_{t',k',t,k}(y_t^k|X) &= \sum_{y_{t'}^{k'}} \varphi_{t,k,t',k'}(y_t^k, y_{t'}^{k'}|X) Q(y_{t'}^{k'}) \\ &\approx \begin{cases} \sum_{y_{t'}^{k'}} \left\langle g_\theta^{kk'}(X_t^k) \otimes h_\theta^{kk'}(X_t^k) \right\rangle_{y_t^k, y_{t'}^{k'}} Q(y_{t'}^{k'}) & t = t' \\ \sum_{y_{t'}^{k'}} K_\sigma\left(t, t'\right) \left\langle r_\theta^{kk'}(X_t^k) \otimes s_\theta^{kk'}(X_t^k) \right\rangle_{y_t^k, y_{t'}^{k'}} Q(y_{t'}^{k'}) & t \neq t' \end{cases}. \end{aligned}$$

The equation can be reformulated in matrix form:

$$-\log \mathbf{m}_{t',k',t,k} \approx \text{Query} \cdot \text{Key}^\top \cdot \text{Value},$$

where

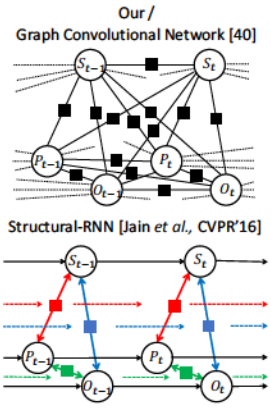
$$\begin{aligned} \langle \mathbf{m}_{t',k',t,k} \rangle_{y_t^k} &= m_{t',k',t,k}(y_t^k|X) \\ \langle \text{Value} \rangle_{y_{t'}^{k'}} &= \begin{cases} Q(y_{t'}^{k'}) & t = t' \\ K_\sigma\left(t, t'\right) \cdot Q(y_{t'}^{k'}) & t \neq t' \end{cases} \\ \text{Query} &= \begin{cases} g_\theta^{kk'}(X_t^k) & t = t' \\ r_\theta^{kk'}(X_t^k) & t \neq t' \end{cases} \\ \text{Key} &= \begin{cases} h_\theta^{kk'}(X_t^k) & t = t' \\ s_\theta^{kk'}(X_t^k) & t \neq t' \end{cases}. \end{aligned}$$

We now link this message form with self attention in Machine Translation [10] and Non-Local Mean in Computer Vision [1]. Self-Attention is expressed as the form of

$$\text{softmax}\Big(\text{Query} \cdot \text{Key}^\top\Big) \cdot \text{Value}$$

with Query, Key, and Value depending on input (termed observation in our case).

In both Self Attention and our message form, the attended weights for Value is dependent on observation. The difference is that we do not have a row-wise softmax activation to make the attended weights sum to 1. The derivation is also similar to Non-Local Means [1]. Note that Machine Translation [10] focuses on the updates for features across temporal regions, Non-Local Mean [1] focuses on the updates for the features across spatial regions, while ours focuses on the updates for the entities prediction (i.e., as a message passing).



Method	Relationship Detection relationship			Relationship Tagging relationship			Relationship Recognition			
	R@50	R@100	mAP	P@1	P@5	P@10	*A* Acc@1	*B* Acc@1	*C* Acc@1	relationship Acc@1
Standard Evaluation for ImageNet Video dataset: A = subject, B = predicate, C = object										
Structural RNN [3]	6.89	8.62	6.89	46.50	33.30	26.94	88.73	27.47	88.52	23.80
Graph Convolution [11]	6.02	7.53	8.21	38.50	30.20	22.70	86.29	24.22	85.77	19.18
GSTEG (Ours)	**7.05**	**8.67**	**9.52**	**51.50**	**39.50**	**28.23**	**90.60**	**28.78**	**89.79**	**25.01**
Zero-Shot Evaluation for ImageNet Video dataset: A = subject, B = predicate, C = object										
Structural RNN [3]	0.12	0.19	0.10	1.36	**1.92**	1.85	70.60	6.71	67.59	2.78
Graph Convolution [11]	**0.16**	0.16	**0.20**	1.37	**1.92**	1.51	75.00	5.32	72.45	3.94
GSTEG (Ours)	**1.16**	**2.08**	0.15	**2.74**	**1.92**	**1.92**	**82.18**	**7.87**	**79.40**	**6.02**
Standard Evaluation for Charades dataset: A = object, B = verb, C = scene										
Structural RNN [3]	23.63	31.15	8.73	17.18	12.24	9.18	42.73	64.32	34.40	12.40
Graph Convolution [11]	23.53	31.10	8.56	16.96	12.23	9.43	42.19	**64.82**	36.11	12.75
GSTEG (Ours)	**24.95**	**33.37**	**9.86**	**19.16**	**12.93**	**9.55**	**43.53**	**64.82**	**40.11**	**14.73**

	Gated Spatio-Temporal Energy Graph (ours)	Graph Convolutional Network [40]	Structural-RNN [Jain *et al.*, CVPR'16]
Fully-Connected / Probabilistic / Length of the Graph	Yes / Yes / Entire Video	Yes / No / Partial Video (~32 frames)	No / No / Entire Video
$\aleph(A)$: set of nodes that contribute to A	All Nodes in Entire Video	All Nodes in Partial Video	Partial Past Nodes in Entire Video (no Future Nodes)
$m_{B\to A}$: message from node B to node A $X_A \in \mathbb{R}^d$: feature of node A, $y_A \in \mathbb{R}^{\lvert\mathcal{Y}_A\rvert}$: its corresponding label	matrix: label compatibility matrix $m_{B\to A} = \boxed{(FC_1(X_A) \otimes FC_2(X_A))}\, y_B$	scalar: similarity between X_A and X_B $m_{B\to A} = \boxed{(FC_1(X_A)^T FC_2(X_B))}\, X_B$	$m_{B\to A} = RNN_1([X_A, X_B])$
y_A: label for node A $FC_\cdot$: fully-connected layer, $RNN_\cdot$: recurrent layer, $[\cdots]$: concatenation	$y_A = softmax\left(FC_3(X_A) + \sum_{B\in\aleph(A)} m_{B\to A}\right)$	$y_A = softmax\left(FC_3\left(m_{A\to A} + \sum_{B\in\aleph(A)} m_{B\to A}\right)\right)$	$y_A = softmax\left(RNN_2\left([m_{B_1\to A}, \cdots, m_{B_{\lvert\aleph(A)\rvert}\to A}, X_A]\right)\right)$

Figure 1. [Bottom] Table summarizing the novelty of our proposed approach v.s. competing methods, [Top-left] Comparison of the graphical structures, [Top-right] Empirical comparisons between our approach and other Structural RNN [3] and Graph Convolution [11]. Our model performs well across all the three tasks.

3. Comparisons with SRNN [3] & GCN [11]

Here, we provide comparisons with Structural-RNN (SRNN) [3] and Graph Convolutional Network (GCN) [11] for comparisons. We note that these approaches are designed for video activity recognition, which cannot be directly applied in video visual relationship detection. In Fig. 1 (top-left), we show how we minimally modifying SRNN and GCN for evaluating them in video relationship detection. The main differences are: 1) our model constructs a fully-connected graph for entire video, while SRNN has a non-fully connected graph and GCN considers only building a graph on partial video (~32 frames), and 2) the message passing across node represents prediction's dependency for our model, while it indicates temporally evolving edge features for SRNN and similarity-reweighted features for GCN.

4. Activity Recognition in Charades

Sigurdsson *et al.* [9] proposed Asynchronous Temporal Fields (AsyncTF) for recognizing 157 video activities. As discussed in Related Work (see Sec. 2), video activity recognition is a downstream task of visual relationship learning: in Charades, each activity (in 157 activities) is a combination of one category *object* and one category in *verb*. We now cast how our model be transformed into video activity recognition. First, we change the output sequence to be $Y = \{Y_t\}_{t=1}^T$, where Y_t is the prediction of video activity. Then, we apply our Gated Spatio-Temporal Energy Graph on top of the sequence of activity predictions. In this design, we achieve the mAP of 33.3%. AsyncTF reported the mAP 18.3% for using only RGB values from a video.

5. Feature Representation in Pre-Reasoning Modules

ImageNet Video. We now provide the details for representing X_t^p, which is the predicate feature in t_{th} chunk of the input instance. Note that we use the relation feature from prior work [8] (the feature can be downloaded form [7]) as our predicate feature. The feature comprises three components: (1) improved dense trajectory (iDT) features from subject trajectory, (2) improved dense trajectory (iDT) features from object trajectory, and (3) relative features describing the relative positions, size, and motion between subject and object trajectories. iDT features are able to capture the movement and also the low-level visual characteristics for an object moving in a short clip. The relative features are able to represent the relative spatio-temporal differences between subject trajectory and object trajectory. Next, the features are post-processed as bag-of-words features after applying a dictionary learning on the original features. Last, three sub-features are concatenated together for representing our predicate feature.

Charades. We use the output feature layer from I3D network [2] to represent our object (X_t^o), verb (X_t^v), and scene feature (X_t^s). The I3D network is pre-trained from Kinetics dataset [2] (the model can be downloaded from [12]) and the output feature layer is the layer before output logits.

6. Intractable Inference during Evaluation

In ImageNet Video dataset, during evaluation, for relation detection and tagging, we have to enumerate all the possible associations of subject or object tracklets. The number of possible associations grows exponentially by the factor of the number of chunks in a video, which will easily become computationally intractable. Note that the problem exists only during evaluation since the ground truth associations (for subject and object tracklets) are given during training. To overcome the issue, we apply the greedy association algorithm described in [8] for efficiently associating subject or object tracklets. The idea is as follows. First, we adopt the inference only in a chunk. Since the message does not pass across chunks, at this step, we don't need to consider associations (for subject or object tracklets) across chunks. In a chunk, for a pair of subject and object tracklet, we have a predicted relation triplet. Then, from two overlapping chunks, we only associate the pair of the subject and object tracklets with the same predicted relation triplet and high tracklets vIoU (i.e., > 0.5). Comparing to the original inference, this algorithm exponentially accelerates the time computational complexity. On the other hand, in Charades, we do not need associate object tracklets. Thus, the intractable computation complexity issue does not exist. The greedy associate algorithm is not required for Charades.

7. Training and Parametrization Details

We specify the training and parametrization details as follows.

ImageNet Video. Throughout all the experiments, we choose Adam [4] with learning rate 0.001 as our optimizer, 32 as our batch size, 30 as the number of training epoch, and 3 as the number of message passing. We initialize the marginals to be the marginals estimated from unary energy.

- Rank number r: 5
- $g_\theta^{kk'}(X_t^k)$: $|X_t^k| \times (|Y_t^k| \times r)$ fully-connected layer, resize to $|Y_t^k| \times r$
- $h_\theta^{kk'}(X_t^k)$: $|X_t^k| \times \times (|Y_{t'}^{k'}| \times r)$ fully-connected layer, resize to $|Y_{t'}^{k'}| \times r$
- $r_\theta^{kk'}(X_t^k)$: $|X_t^k| \times 1024$ fully-connected layer, ReLU Activation, Dropout with rate 0.3, 1024×1024 fully-connected layer, ReLU Activation, Dropout with rate 0.3, $1024 \times (|Y_t^k| \times r)$ fully-connected layer, resize to $|Y_t^k| \times r$
- $s_\theta^{kk'}(X_t^k)$: $|X_t^k| \times 1024$ fully-connected layer, ReLU Activation, Dropout with rate 0.3, 1024×1024 fully-connected layer, ReLU Activation, Dropout with rate 0.3, $1024 \times (|Y_{t'}^{k'}| \times r)$ fully-connected layer, resize to $|Y_{t'}^{k'}| \times r$
- σ: 10
- $w_\theta^{kk'}(X_t^k)$: $|X_t^k| \times |Y_t^k|$ fully-connected layer

Charades: Throughout all the experiments, we choose SGD with learning rate 0.005 as our optimizer, 40 as our batch size, 5 as the number of training epoch, and 5 as the number of message passing. We initialize the marginals to be the marginals estimated from unary energy.

- Rank number r: 5
- $g_\theta^{kk'}(X_t^k)$: $|X_t^k| \times (|Y_t^k| \times r)$ fully-connected layer, resize to $|Y_t^k| \times r$
- $h_\theta^{kk'}(X_t^k)$: $|X_t^k| \times (|Y_{t'}^{k'}| \times r)$ fully-connected layer, resize to $|Y_{t'}^{k'}| \times r$
- $r_\theta^{kk'}(X_t^k)$: $|X_t^k| \times (|Y_t^k| \times r)$ fully-connected layer, resize to $|Y_t^k| \times r$
- $s_\theta^{kk'}(X_t^k)$: $|X_t^k| \times \times (|Y_{t'}^{k'}| \times r)$ fully-connected layer, resize to $|Y_{t'}^{k'}| \times r$
- σ: 300
- $w_\theta^{kk'}(X_t^k)$: $|X_t^k| \times |Y_t^k|$ fully-connected layer

8. Parametrization in Leveraging Language Priors

Additional networks in the experiments towards leveraging language priors are parametrized as follows:

- d: 300 (because we use 300-dim. Glove [6] features)
- $u_\theta(\cdot)$: $d \times 1024$ fully-connected layer, ReLU Activation, Dropout with rate 0.3, 1024×1024 fully-connected layer, ReLU Activation, Dropout with rate 0.3, 1024×1 fully-connected layer
- $v_\theta(\cdot)$: $d \times 1024$ fully-connected layer, ReLU Activation, Dropout with rate 0.3, 1024×1024 fully-connected layer, ReLU Activation, Dropout with rate 0.3, 1024×1 fully-connected layer

9. Category Set in Dataset

For clarity, we use bullet points for referring to the category choice in datasets for the different entity.

- *subject / object* in ImageNet Video (total 35 categories)
  - airplane, antelope, ball, bear, bicycle, bird, bus, car, cat, cattle, dog, elephant, fox, frisbee, giant panda, hamster, horse, lion, lizard, monkey, motorcycle, person, rabbit, red panda, sheep, skateboard, snake, sofa, squirrel, tiger, train, turtle, watercraft, whale, zebra

- *predicate* in ImageNet Video (total 132 categories)
  - taller, swim behind, walk away, fly behind, creep behind, lie with, move left, stand next to, touch, follow, move away, lie next to, walk with, move next to, creep above, stand above, fall off, run with, swim front, walk next to, kick, stand left, creep right, sit above, watch, swim with, fly away, creep beneath, front, run past, jump right, fly toward, stop beneath, stand inside, creep left, run next to, beneath, stop left, right, jump front, jump beneath, past, jump toward, sit front, sit inside, walk beneath, run away, stop right, run above, walk right, away, move right, fly right, behind, sit right, above, run front, run toward, jump past, stand with, sit left, jump above, move with, swim beneath, stand behind, larger, walk past, stop front, run right, creep away, move toward, feed, run left, lie beneath, fly front, walk behind, stand beneath, fly above, bite, fly next to, stop next to, fight, walk above, jump behind, fly with, sit beneath, sit next to, jump next to, run behind, move behind, swim right, swim next to, hold, move past, pull, stand front, walk left, lie above, ride, next to, move beneath, lie behind, toward, jump left, stop above, creep toward, lie left, fly left, stop with, walk toward, stand right, chase, creep next to, fly past, move front, run beneath, creep front, creep past, play, lie inside, stop behind, move above, sit behind, faster, lie right, walk front, drive, swim left, jump away, jump with, lie front, left
- *verb* in Charades (total 33 categories)
  - awaken, close, cook, dress, drink, eat, fix, grasp, hold, laugh, lie, make, open, photograph, play, pour, put, run, sit, smile, sneeze, snuggle, stand, take, talk, throw, tidy, turn, undress, walk, wash, watch, work
- *object* in Charades (total 38 categories)
  - None, bag, bed, blanket, book, box, broom, chair, closet/cabinet, clothes, cup/glass/bottle, dish, door, doorknob, doorway, floor, food, groceries, hair, hands, laptop, light, medicine, mirror, paper/notebook, phone/camera, picture, pillow, refrigerator, sandwich, shelf, shoe, sofa/couch, table, television, towel, vacuum, window
- *scene* in Charades (total 16 categories)
  - Basement, Bathroom, Bedroom, Closet / Walk-in closet / Spear closet, Dining room, Entryway, Garage, Hallway, Home Office / Study, Kitchen, Laundry room, Living room, Other, Pantry, Recreation room / Man cave, Stairs

References

[1] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 60–65. IEEE, 2005. 1

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4724–4733. IEEE, 2017. 2

[3] 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. 2

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[5] 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. 1

[6] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 1, 3

[7] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. *URL for ImageNet Video*. https://lms.comp.nus.edu.sg/research/VidVRD.html. 2

[8] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1300–1308. ACM, 2017. 2, 3

[9] Gunnar A Sigurdsson, Santosh Kumar Divvala, Ali Farhadi, and Abhinav Gupta. Asynchronous temporal fields for action recognition. In *CVPR*, volume 5, page 7, 2017. 2

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 1

[11] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. *arXiv preprint arXiv:1806.01810*, 2018. 2

[12] Brian Zhang, Joao Carreira, Viorica Patraucean, Diego de Las Casas, Chloe Hillier, and Andrew Zisserman. *URL for I3D*. https://github.com/deepmind/kinetics-i3d. 2