Detecting Visual Relationships with Deep Relational Networks

Bo Dai Yuqi Zhang Dahua Lin
Department of Information Engineering, The Chinese University of Hong Kong
db014@ie.cuhk.edu.hk zy016@ie.cuhk.edu.hk dhlin@ie.cuhk.edu.hk

Abstract

Relationships among objects play a crucial role in image understanding. Despite the great success of deep learning techniques in recognizing individual objects, reasoning about the relationships among objects remains a challenging task. Previous methods often treat this as a classification problem, considering each type of relationship (e.g. "ride") or each distinct visual phrase (e.g. "person-ride-horse") as a category. Such approaches are faced with significant difficulties caused by the high diversity of visual appearance for each kind of relationships or the large number of distinct visual phrases. We propose an integrated framework to tackle this problem. At the heart of this framework is the Deep Relational Network, a novel formulation designed specifically for exploiting the statistical dependencies between objects and their relationships. On two large data sets, the proposed method achieves substantial improvement over state-of-the-art.

1. Introduction

Images in the real world often involve multiple objects that interact with each other. To understand such images, being able to recognize individual objects is generally not sufficient. The *relationships* among them also contain crucial messages. For example, image captioning, a popular application in computer vision, can generate richer captions based on relationships in addition to objects in the images. Thanks to the advances in deep learning, the past several years witness remarkable progress in several key tasks in computer vision, such as *object recognition* [2], *scene classification* [3], and *attribute detection* [4]. However, visual relationship detection remains a *very difficult* task. On Visual Genome [5], a large dataset designed for structural image understanding, the state-of-the-art can only obtain 11.79% of Recall@50 [1]. This performance is clearly far from being satisfactory.

A natural approach to this problem is to treat it as a classification task. Early attempts [6] used to consider different combinations of objects and relationship predicates (known as *visual phrases*) as different classes. While it may work in a restricted context where the number of possible combinations is moderate, such strategy would be met with a fundamental difficulty in general – an extremely large number of imbalanced classes. As a case in point, Visual Genome [5] contains over $75K$ distinct visual phrases, and the number of samples for each phrase ranges from just a handful to over $10K$. Even the most sophisticated classifier would suffer facing such a large and highly imbalanced class space.

An alternative strategy is to consider each type of re-

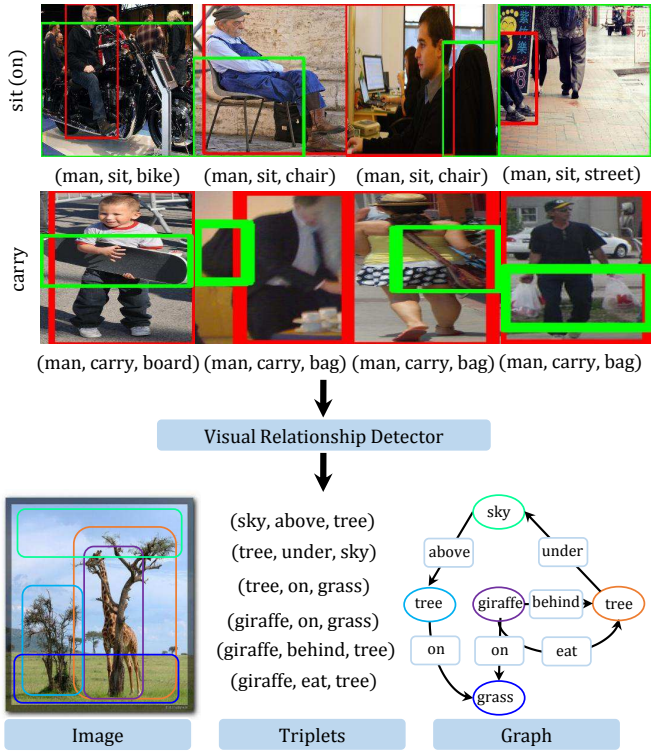


Figure 1: Visual relationships widely exist in real-world images. Here are some examples from the VRD [1] dataset, with relationship predicates *"sit"* and *"carry"*. We develop a method that can effectively detect such relationships from a given image. On top of that, a scene graph can be constructed.

lationship predicates as a class. Whereas the number of classes is drastically smaller, along with this change also comes with an undesirable implication, namely the substantially increased diversity within each class. To be more specific, phrases with different object categories are considered to be in the same class, as long as they have the same type of relationship predicates. Consequently, the images in each class are highly diverse – some images in the same class may even share nothing in common, *e.g. "mountain-near-river"* and *"person-near-dog"*. See Figure 1 for an illustration. Our experiments suggest that even with the model capacity of deep networks, handling the intra-class diversity at this level remains very difficult.

In this work, we develop a new framework to tackle the problem of *visual relationship detection*. This framework formulates the prediction output as a triplet in the form of *(subject, predicate, object)*, and jointly infers their class labels by exploiting two kinds of relations among them, namely *spatial configuration* and *statistical dependency*. Such relations are ubiquitous, informative, and more importantly they are often more reliable than visual appearance.

It is worth emphasizing that the formulation of the proposed model is significantly different from previous relational models such as conditional random fields (CRFs) [7]. Particularly, in our formulation, the statistical inference procedure is embedded into a deep neural network called *Deep Relational Network (DR-Net)* via iteration unrolling. The formulation of DR-Net moves beyond the conventional scope, extending the expressive power of Deep Neural Networks (DNNs) to relational modeling. This new way of formulation also allows the model parameters to be learned in a discriminative fashion, using the latest techniques in deep learning. On two large datasets, the proposed framework outperforms not only the classification-based methods but also the CRFs based on deep potentials.

To sum up, the major contributions of this work consist in two aspects: (1) DR-Net, a novel formulation that combines the strengths of statistical models and deep learning; and (2) an effective framework for visual relationship detection[1], which brings the state-of-the-art to a new level.

2. Related Work

Over the past decade, there have been a number of studies that explore the use of *visual relationships*. Earlier efforts often focus on *specific* types of relationships, such as positional relations [8–12] and actions (*i.e.* interactions between objects) [13–23]. In most of these studies, relationships are usually extracted using simple heuristics or handcrafted features, and used as an auxiliary components to facilitate other tasks, such as object recognition [24–32], image classification and retrieval [33,34], scene understanding and generation [35–41], as well as text grounding [42–44]. They are essentially different from our work, which aims to provide a method dedicated to *generic* visual relationship detection. On a unified framework, our method can recognize a wide variety of relationships, such as relative positions (*"behind"*), actions (*"eat"*), functionals (*"part of"*), and comparisons (*"taller than"*).

Recent years have seen new methods developed specifically for detecting visual relationships. An important family of methods [6,45,46] consider each distinct combination of object categories and relationship predicates as a distinct class (often referred to as a *visual phrase*). Such methods would face difficulties in a general context, where the number of such combinations can be very large. An alternative paradigm that considers relationship predicates and object categories separately becomes more popular in recent efforts. Vedantam *et al*. [47] presented a study along this line using synthetic clip-arts. This work, however, relies on multiple synthetic attributes that are difficult to obtain from natural images. Fang *et al*. [48] proposed to incorporate relationships in an image captioning framework. This work treats object categories and relationship predicates uniformly as words, and does not discuss how to tackle the various challenges in relationship detection.

The method proposed recently by Lu *et al*. [1] is the most related. In this method, pairs of detected objects are fed to a classifier, which combines appearance features and a language prior for relationship recognition. Our method differs in two aspects: (1) We exploit both spatial configurations and statistical dependencies among *relationship predicates*, *subjects*, and *objects*, via a Deep Relational Network, instead of simply fusing them as different features. (2) Our framework, from representation learning to relational modeling, is integrated into a single network that is learned in an end-to-end fashion. Experiments show that the proposed framework performs substantially better in all different task settings. For example, on two large datasets, the *Recall@50* of relationship predicate recognition are respectively raised from 47.9% to 80.8% and from 53.5% to 88.3%.

3. Visual Relationship Detection

Visual relationships play a crucial role in image understanding. Whereas a relationship may involve multiple parties in general, many important relationships, including *relative positions* (*e.g. "above"*) and *actions* (*e.g. "ride"*) occur between exactly two objects. In this paper, we focus on such relationships. In particular, we follow a widely adopted convention [1, 6] and characterize each visual relationship by a *triplet* in the form of (s, r, o), *e.g. (girl, on, horse)* and *(man, eat, apple)*. Here, s, r, and o respectively denote the *subject category*, the *relationship predicate*, and the *object category*. The task is to locate all visual relationships from a given image, and infer the triplets.

[1] code available at github.com/doubledaibo/drnet

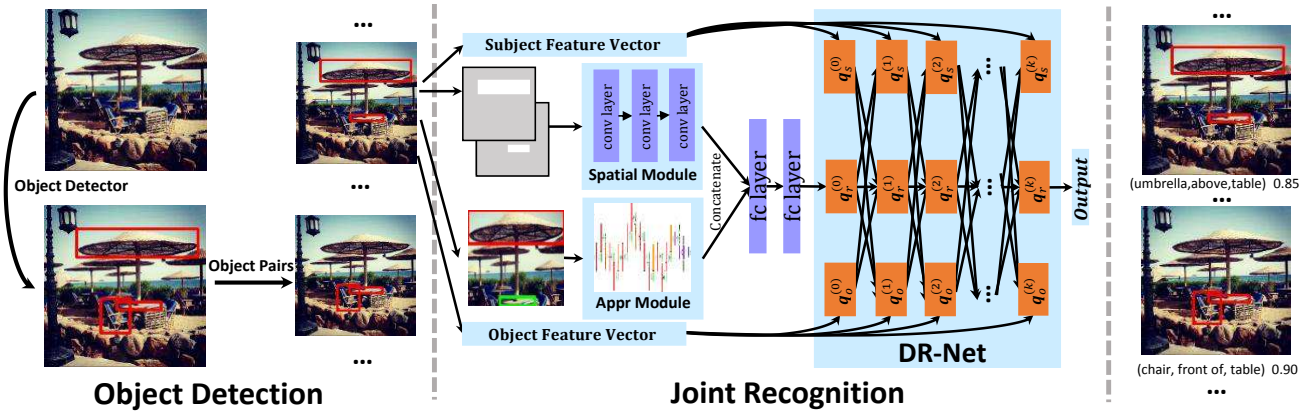


Figure 2: The proposed framework for visual relationship detection. Given an image, it first employs an object detector to locate individual objects. Each object also comes with an appearance feature. For each pair of objects, the corresponding local regions and the spatial masks will be extracted, which, together with the appearance features of individual objects, will be fed to the DR-Net. The DR-Net will jointly analyze all aspects and output $\mathbf{q}_s$, $\mathbf{q}_r$, and $\mathbf{q}_o$, the predicted category probabilities for each component of the triplet. Finally, the triplet (s, r, o) will be derived by choosing the most probable categories for each component.

3.1. Overall Pipeline

As mentioned, there are two different paradigms for relationship detection: one is to consider each distinct triplet as a different category (also known as *visual phrases* [6]), the other is to recognize each component individually. The former is not particularly suitable for generic applications, due to difficulties like the excessively large number of classes and the imbalance among them. In this work, we adopt the latter paradigm and aim to take its performance to a next level. Particularly, we focus on developing a new method that can effectively capture the rich relations (both *spatial* and *semantic*) among the three components in a triplet and exploit them to improve the prediction accuracy.

As shown in Figure 2, the overall pipeline of our framework comprises three stages, as described below.

(1) Object detection. Given an image, we use an object detector to locate a set of candidate objects. In this work, we use Faster RCNN [2] for this purpose. Each candidate object comes with a bounding box and an appearance feature, which will be used in the joint recognition stage for predicting the object category.

(2) Pair filtering. The next step is to produce a set of *object pairs* from the detected objects. With n detected objects, we can form $n(n-1)$ pairs. We found that a considerable portion of these pairs are *obviously* meaningless and it is unlikely to recognize important relationships therefrom. Hence, we introduce a low-cost neural network to filter out such pairs, so as to reduce the computational cost of the next stage. This filter takes into account both the spatial configurations (*e.g.* objects too far away are unlikely to be related) and object categories (*e.g.* certain objects are unlikely to form a meaningful relationship).

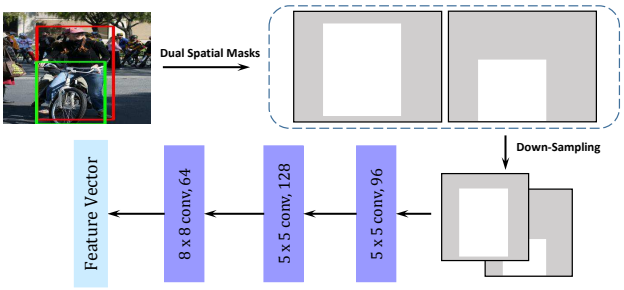


Figure 3: This figure illustrates the process of spatial feature vector generation. The structure of our spatial module is also presented in this figure.

(3) Joint recognition. Each retained pair of objects will be fed to the *joint recognition* module. Taking into account multiple factors and their relations, this module will produce a triplet as the output.

3.2. Joint Recognition

In joint recognition, multiple factors are taken into consideration. These factors are presented in detail below.

(1) Appearance. As mentioned, each detected object comes with an appearance feature, which can be used to infer its category. In addition, the type of the relationship may also be reflected in an image visually. To utilize this information, we extract an appearance feature for each *candidate pair* of objects, by applying a CNN [49, 50] to an *enclosing box*, *i.e.* a bounding box that encompasses both objects with a small margin. The appearance inside the enclosing box captures not only the objects themselves but also the surrounding context, which is often useful when reasoning

about the relationships.

(2) Spatial Configurations. The relationship between two objects is also reflected by the spatial configurations between them, *e.g*. their relative positions and relative sizes. Such cues are complementary to the appearance of individual objects, and resilient to photometric variations, *e.g*. the changes in illumination.

To leverage the spatial configurations, we are facing a question: *how to represent it in a computer?* Previous work [9] suggests a list of geometric measurements. While simple, this way may risk missing certain aspects of the configurations. In this work, we instead use *dual spatial masks* as the representation, which comprise two binary masks, one for the subject and the other for the object. The masks are derived from the bounding boxes and may overlap with each other, as shown in Figure 3. The masks are downsampled to the size 32×32, which we found empirically is a good balance between fidelity and cost. (We have tried mask sizes of 8, 16, 32, 64 and 128, resulting top-1 recalls are 0.47, 0.48, 0.50, 0.51 and 0.51.) The dual spatial masks for each candidate pair will be compressed into a 64-dimensional vector via three convolutional layers.

(3) Statistical Relations. In a triplet (s, r, o), there exist strong statistical dependencies between the relationship predicate r and the object categories s and o. For example, *(cat, eat, fish)* is common, while *(fish, eat, cat)* or *(cat, ride, fish)* is very unlikely. On Visual Genome, the entropy of the prior distribution $p(r)$ is 2.88, while that of the conditional distribution $p(r|s, o)$ is 1.21. This difference is a clear evidence of the statistical dependency.

To exploit the statistical relations, we propose *Deep Relational Network (DR-Net)*, a novel formulation that incorporates statistical relational modeling into a deep neural network framework. In our experiments, we found that the use of such relations can effectively resolve the ambiguities caused by visual or spatial cues, thus substantially improving the prediction accuracy.

(4) Integrated Prediction. Next, we describe how these factors are actually combined. As shown in Figure 2, for each candidate pair, the framework extracts the appearance feature and the spatial feature, respectively via the appearance module and the spatial module. These two features are subsequently concatenated and further compressed via two fully-connected layers. This *compressed pair feature*, together with the appearance features of individual objects will be fed to the DR-Net for joint inference. Through multiple inference units, whose parameters capture the statistical relations among triplet components, the *DR-Net* will output the posterior probabilities of s, r, and o. Finally, the framework produces the prediction by choosing the most probable classes for each of these components.

In the training, all stages in our framework, namely object detection, pair filtering and joint recognition are trained respectively. As for joint recognition, different factors will be integrated into a single network and jointly fine-tuned to maximize the joint probability of the ground-truth triplets.

4. Deep Relational Network

As shown above, there exist strong statistical relations among the object categories s and o and the relationship predicates r. Hence, to accurately recognize visual relationships, it is important to exploit such information, especially when the visual cues are ambiguous.

4.1. Revisit of CRF

The *Conditional Random Field (CRF)* [7] is a classical formulation to incorporate statistical relations into a discriminative task. Specifically, for the task of recognizing visual relationships, the CRF can be formulated as

$$p(r, s, o|\mathbf{x}_r, \mathbf{x}_s, \mathbf{x}_o) = \frac{1}{Z} \exp\left(\Phi(r, s, o|\mathbf{x}_r, \mathbf{x}_s, \mathbf{x}_o; \mathbf{W})\right). \quad (1)$$

Here, $\mathbf{x}_r$ is the *compressed pair feature* that combines both the appearance of the enclosing box and the spatial configurations; $\mathbf{x}_s$ and $\mathbf{x}_o$ are the appearance features respectively for the subject and the object; $\mathbf{W}$ denotes the model parameters; and Z is the normalizing constant, whose value depends on the parameters $\mathbf{W}$. The joint potential Φ can be expressed as a sum of individual potentials as

$$\begin{aligned}\Phi = &\psi_a(s|\mathbf{x}_s; \mathbf{W}_a) + \psi_a(o|\mathbf{x}_o; \mathbf{W}_a) + \psi_r(r|\mathbf{x}_r; \mathbf{W}_r) \\ &+ \varphi_{rs}(r, s|\mathbf{W}_{rs}) + \varphi_{ro}(r, o|\mathbf{W}_{ro}) + \varphi_{so}(s, o|\mathbf{W}_{so}). \quad (2)\end{aligned}$$

Here, the unary potential ψ_a associates individual objects with their appearance; ψ_r associates the relationship predicate with the feature $\mathbf{x}_r$; while the binary potentials φ_{rs}, φ_{ro} and φ_{so} capture the statistical relations among the relationship predicate r, the subject category s, and the object category o.

CRF formulations like this have seen wide adoption in computer vision literatures [51, 52] over the past decade, and have been shown to be a viable way to capture statistical dependencies. However, the success of CRF is limited by several issues: First, learning CRF requires computing the normalizing constant Z, which can be very expensive and even intractable, especially when cycles exist in the underlying graph, like the formulation above. Hence, approximations are often used to circumvent this problem, but they sometimes result in poor estimates. Second, when cyclic dependencies are present, variational inference schemes such as mean-field methods [53] and loopy belief propagation [54], are widely used to simplify the computation. This often leaves a gap between the objective of inference and that of training, thus leading to suboptimal results.

4.2. From CRF to DR-Net

Inspired by the success of deep neural networks [49, 50], we explore an alternative approach to relational modeling, that is, to *unroll* the inference into a feed-forward network.

Consider the CRF formulated above. Given s and o, then the posterior distribution of r is given by

$$p(r|s,o,\mathbf{x}_r;\mathbf{W}) \propto \exp\left(\psi_r(r|\mathbf{x}_r;\mathbf{W}_r)+ \varphi_{rs}(r,s|\mathbf{W}_{rs})+\varphi_{ro}(r,o|\mathbf{W}_{ro})\right). \quad (3)$$

In typical formulations, $\psi_r(r|\mathbf{x}_r)$ is often devised to be a linear functional of $\mathbf{x}_r$ for each r. Let $\mathbf{W}_{rs}$ and $\mathbf{W}_{ro}$ be matrices such that $\mathbf{W}_{rs}(r,s)=\varphi_{rs}(r,s)$ and $\mathbf{W}_{ro}(r,o)=\varphi_{ro}(r,o)$, and let $\mathbf{q}_r$ be a vector of the posterior probabilities for r, then the formula above can be rewritten as[2]

$$\mathbf{q}_r = \boldsymbol{\sigma}\left(\mathbf{W}_r\mathbf{x}_r+\mathbf{W}_{rs}\mathbf{1}_s+\mathbf{W}_{ro}\mathbf{1}_o\right). \quad (4)$$

Here, $\boldsymbol{\sigma}$ denotes the *softmax* function. $\mathbf{1}_s$ and $\mathbf{1}_o$ are one-hot indicator vectors for s and o. It can be shown that this is the optima to the optimization problem below:

$$\max_{\mathbf{q}} \; E_q\left[\psi_r(r|\mathbf{x}_r;\mathbf{W}_r)+ \varphi_{rs}(r,s|\mathbf{W}_{rs})+\varphi_{ro}(r,o|\mathbf{W}_{ro})\right]+H_q(\mathbf{q}). \quad (5)$$

Based on this optimization problem, the solution given in Eq.(4) can be generalized to the case where s and o are not deterministic and the knowledge of them are instead given by probabilistic vectors $\mathbf{q}_s$ and $\mathbf{q}_o$, as follows:

$$\mathbf{q}_r = \boldsymbol{\sigma}\left(\mathbf{W}_r\mathbf{x}_r+\mathbf{W}_{rs}\mathbf{q}_s+\mathbf{W}_{ro}\mathbf{q}_o\right). \quad (6)$$

Similar derivation also applies to the inference of s and o conditioned on other components. Together, we can obtain a set of *updating formulas* as below:

$$\begin{aligned}\mathbf{q}'_s &= \boldsymbol{\sigma}\left(\mathbf{W}_a\mathbf{x}_s+\mathbf{W}_{sr}\mathbf{q}_r+\mathbf{W}_{so}\mathbf{q}_o\right),\\ \mathbf{q}'_r &= \boldsymbol{\sigma}\left(\mathbf{W}_r\mathbf{x}_r+\mathbf{W}_{rs}\mathbf{q}_s+\mathbf{W}_{ro}\mathbf{q}_o\right),\\ \mathbf{q}'_o &= \boldsymbol{\sigma}\left(\mathbf{W}_a\mathbf{x}_o+\mathbf{W}_{os}\mathbf{q}_s+\mathbf{W}_{or}\mathbf{q}_r\right). \end{aligned} \quad (7)$$

These formulas take the current probability vectors $\mathbf{q}_s$, $\mathbf{q}_r$, and $\mathbf{q}_o$ as inputs, and output the updated versions $\mathbf{q}'_s$, $\mathbf{q}'_r$ and $\mathbf{q}'_o$. From the perspective of neural networks, these formulas can also be viewed as a *computing layer*. In this sense, the iterative updating procedure can be *unrolled* into a network that comprises a sequence of such layers. We call this network the *Deep Relational Network (DR-Net)*, as it relates multiple variables, and refer to its building blocks, *i.e.* the computing layers mentioned above, as *inference units*.

Discussion DR-Net is for relational modeling, which is different from those methods for feature/modality combination. Specifically, *object categories* and *relationship predicates* are two distinct domains that are statistically related. The former is not an extra feature of the latter; while the latter is not a feature of the former either. DR-Net captures the relations between them via the links in the inference units, rather than combining them using a fusion layer.

The basic formulation in Eq.7 comes with several symmetry constraints: $\mathbf{W}_{sr}=\mathbf{W}_{rs}^T$, $\mathbf{W}_{so}=\mathbf{W}_{os}^T$, and $\mathbf{W}_{ro}=\mathbf{W}_{or}^T$. In addition, all inference units share the same set of weights. However, from a pragmatic standpoint, one may also consider lifting these constraints, *e.g.* allowing each inference units to have their own weights. This may potentially increase the expressive power of the network. We will compare these two settings, namely with and without weight sharing, in our experiments.

A DR-Net can also be considered as a special form of the Recurrent Neural Network (RNN) – at each step it takes in a fixed set of inputs, *i.e.* the observed features $\mathbf{x}_s$, $\mathbf{x}_r$, and $\mathbf{x}_o$, and refines the estimates of posterior probabilities.

4.3. Comparison with Other Formulations

There are previous efforts that also explore the incorporation of relational structures with deep networks [51, 55–57]. The deep structured models presented in [55, 56, 58] combine a deep network with an MRF or CRF on top to capture the relational structures among their outputs. In these works, classical message-passing methods are used in training and inference. Zheng *et al*. [51] proposed a framework for image segmentation, which adopts an apparently similar idea, that is, to reformulate a structured model into a neural network by turning inference updates into neural layers. In addition to the fact that this work is in a fundamentally different domain (high-level understanding vs. low-level vision), they focused on capturing dependencies among elements in the same domain, *e.g.* those among pixel-wise labels. From a technical view, DR-Net is more flexible, *e.g.* it can handle graphs with nodes of different cardinalities and edges of different types. In [51], the message passing among pixels is *approximately* instantiated using CNN filters and this is primarily suited for grid structures; while in DR-Net, the inference steps are exactly reproduced using fully-connected layers. Hence, it can be applied to capture relationships of arbitrary structures. SPENs introduced in [57] define a neural network serving as an energy function over observed features for multi-label classification. SPENs are used to measure the consistency of configurations, while DR-Net is used to find a good configuration of variables. Also, no inference unrolling is involved in SPENs learning.

5. Experiments

We tested our model on two datasets: (1) **VRD**: the dataset used in [1], containing $5,000$ images and $37,993$ visual relationship instances that belong to $6,672$ triplet types. We follow the train/test split in [1]. (2) **sVG**: a substantially larger subset constructed from Visual Genome [5]. *sVG* contains $108K$ images and $998K$ relationship instances that

[2] A proof of this statement is provided in the supplemental materials.

		Predicate Recognition		Union Box Detection		Two Boxes Detection	
		Recall@50	Recall@100	Recall@50	Recall@100	Recall@50	Recall@100
VRD	VP [6]	0.97	1.91	0.04	0.07	-	-
	Joint-CNN [48]	1.47	2.03	0.07	0.09	0.07	0.09
	VR [1]	47.87	47.87	16.17	17.03	13.86	14.70
	DR-Net	**80.78**	**81.90**	19.02	22.85	16.94	20.20
	DR-Net + pair filter	-	-	**19.93**	**23.45**	**17.73**	**20.88**
sVG	VP [6]	0.63	0.87	0.01	0.01	-	-
	Joint-CNN [48]	3.06	3.99	1.24	1.60	1.21	1.58
	VR [1]	53.49	54.05	13.80	17.39	11.79	14.84
	DR-Net	**88.26**	**91.26**	20.28	25.74	17.51	22.23
	DR-Net + pair filter	-	-	**23.95**	**27.57**	**20.79**	**23.76**

Table 1: Comparison with baseline methods, using *Recall@50* and *Recall@100* as the metrics. We use "-" to indicate *"not applicable"*. For example, no results are reported for *DR-Net + pair filter* on Predicate Recognition, as in this setting, pairs are given, and thus pair filtering can not be applied. Also, no results are reported for *VP* on Two Boxes detection, as VP detects the entire instance as a single entity.

		A_1	A_2	S	A_1S	A_1SC	A_1SD	A_2SD	A_2SDF
VRD	Predicate Recognition	63.39	65.93	64.72	71.81	72.77	80.66	**80.78**	-
	Union Box Detection	12.01	12.56	13.76	16.04	16.37	18.15	**19.02**	**19.93**
	Two Boxes Detection	10.71	11.22	12.16	14.38	14.66	16.12	**16.94**	**17.73**
sVG	Predicate Recognition	72.13	72.54	75.18	79.10	79.18	88.00	**88.26**	-
	Union Box Detection	13.24	13.84	14.01	16.04	16.08	20.21	**20.28**	**23.95**
	Two Boxes Detection	11.35	11.98	12.07	13.77	13.81	17.42	**17.51**	**20.79**

Table 2: Comparison of different variants of the proposed method, using *Recall@50* as the metric.

belong to $74,361$ triplet types. All instances are randomly partitioned into disjoint training and testing sets, which respectively contain $799K$ and $199K$ instances.

5.1. Experiment Settings

Model training. In all experiments, we trained our model using Caffe [59]. The appearance module is initialized with a model pre-trained on ImageNet, while the spatial module and the DR-Net are initialized randomly. After initialization, the entire network is jointly optimized using SGD.

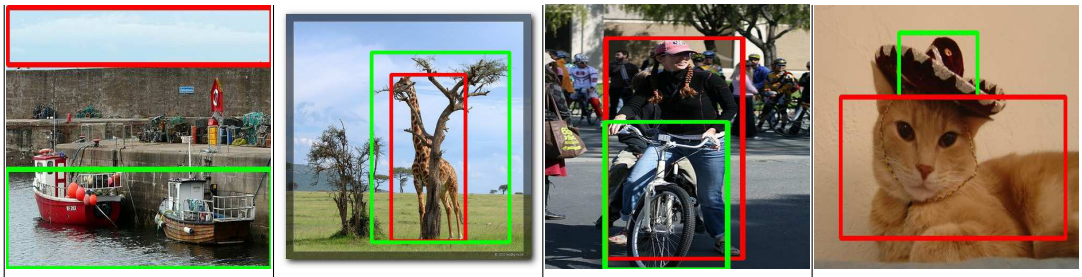
Performance metrics. Following [1], we use *Recall@K* as the major performance metric, which is the the fraction of ground-truth instances that are correctly recalled in top K predictions. Particularly, we report *Recall@100* and *Recall@50* in our experiments. The reason of using *recall* instead of *precision* is that the annotations are incomplete, where some true relationships might be missing.

Task settings. Like in [1], we studied three task settings: **(1) Predicate recognition**: this task focuses on the accuracy of *predicate* recognition, where the labels and the locations of both the *subject* and *object* are given. **(2) Union box detection**: this task treats the whole triplet as a union bounding box. A prediction is considered correct if all three elements in a triplet (s, r, o) are correctly recognized, and the IoU between the predicted box and the ground-truth is above 0.5. **(3) Two boxes detection**: this is similar to the one above, except that it requires the IoU metrics for the subject and the object are both above 0.5. This is relatively more challenging.

5.2. Comparative Results

Compare with baselines. We compared our method with the following methods under all three task settings outlined above. (1) **Visual Phrase(VP)** [6]: a representative approach that treats each distinct triplet as a different class. and employs a DPM detector [60] for each class. (2) **Joint-CNN** [48]: a neural network [49] that has $2N$+K-way outputs, jointly predicts the class responses for subject, object, and relationship predicate. (3) **Visual Relationship (VR)** [1]: This is the state-of-the-art and is the most closely related work.

Table 1 compares the results. On both datasets, we observed: (1) VP [6] performs very poorly, failing in most cases, as it is difficult to cope with such a huge and imbalanced class space. (2) Joint-CNN [48] also works poorly, as it's hard for the CNN to learn a common feature representation for both relationship predicates and objects. (3) VR [1] performs substantially better than the two above. However, the performance remains unsatisfactory. (4) The proposed method outperforms the state-of-the-art method *VR* [1] by a considerable margin in all three tasks. Compared to *VR*, it improves the *Recall@100* of *predicate recognition* by over



VR [1]	(sky, **in**, water)	(giraffe, **have**, tree)	(woman, ride, bicycle)	(cat, **have**, hat)
A_1	(sky, **on**, water)	(giraffe, **have**, tree)	(woman, **behind**, bicycle)	(cat, **on**, hat)
S	(sky, above, water)	(giraffe, **in**, tree)	(woman, **wear**, bicycle)	(cat, **have**, hat)
A_1S	(sky, above, water)	(giraffe, behind, tree)	(woman, **wear**, bicycle)	(cat, **have**, hat)
A_1SC	(sky, above, water)	(giraffe, behind, tree)	(woman, ride, bicycle)	(cat, **have**, hat)
A_1SD	(sky, above, water)	(giraffe, behind, tree)	(woman, ride, bicycle)	(cat, wear, hat)

Table 3: This table lists predicate recognition results for some object pairs. Images containing these pairs are listed in the first row, where the red and green boxes respectively correspond to the subjects and the objects. The most probable predicate predicted by different methods are listed in the following rows, in which **black** indicates wrong prediction and red indicates correct prediction.

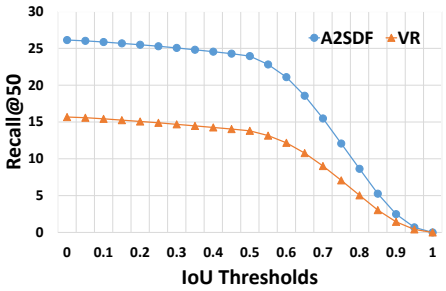


Figure 4: This figure shows the performance on the *union-box detection* task with different IoU thresholds.

30% on both datasets. Thanks to the remarkably improved accuracy in recognizing the relationship predicates, the performance gains on the other two tasks are also significant. (5) Despite the significant gain compared to others, the recalls on *union box detection* and *two boxes detection* remains weak. This is primarily ascribed to the limitations of the object detectors. As shown in Figure 4, we observe that the object detector can only obtain about 30% of object recall, measured by *Recall@50*. To improve on these tasks, a more sophisticated object detector is needed.

Compare different configs. We also compared different variants of the proposed method, in order to identify the contributions of individual components listed below: (1)**Pair (F)ilter**: the pair filter discussed in section 3, used to filter out object pairs with trivial relationships. (2)**(A)ppearance Module**: the appearance module, which has two versions, A_1: based on VGG16 [49], which is also the network used in VR [1], A_2: based on ResNet101 [50]. (3)**(S)patial Module**: the network to capture the spatial configs, as mentioned in section 3. (4)**(C)RF**: a classical CRF formulation, used as a replacement of the DR-Net to capture statistical dependencies. (5)**(D)R-Net**: the DR-Net discussed in section 4. The name of a configuration is the concatenation of abbrevations of involved components, *e.g.*, the configuration named A_1SC contains an appearance module based on VGG16, a spatial module, and a CRF.

In Table 2, we compared A_1, A_2, S, A_1S, A_1SC, A_1SD, A_2SD and A_2SDF. The results show: (1) Using better networks (ResNet-101 vs. VGG16) can moderately improve the performance. However, even with state-of-the-art network A_2, visual relationship detection could not be done effectively using appearance information alone. (2) The combination of appearance and spatial configs considerably outperforms each component alone, suggesting that visual appearances and spatial configurations are complementary to each other. (3) The statistical dependencies are important. However, CRF is not able to effectively exploit them. With the use of DR-Net, the performance gains are significant. We evaluated the perplexities of the predictions for our model *with* and *without* DR-Net, which are 2.64 and 3.08. These results show the benefit of exploiting statistical dependencies for joint recognition.

Table 3 further shows the predicted relationships on several example images. The first two columns show that the incorporation of spatial configuration can help detect positional relationships. The third column shows that the use of statistical dependencies can help to resolve the ambiguities in the relationship predicates. Finally, the fourth column shows that for subtle cases, DR-Net can identify the relationship predicate more accurately than the config that relies on CRF.

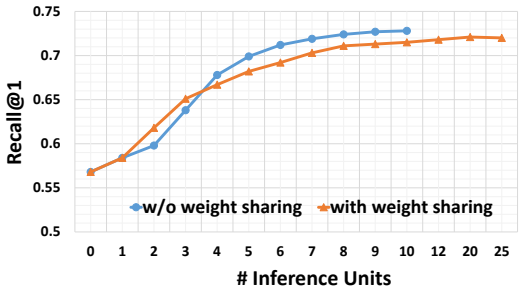


Figure 5: This figure shows the recall curves of two possible settings in DR-Net. In each setting, we change the number of inference units to see how the recall changes.

Average Similarity				
VR [1]	A_1	S	A_1S	A_1SD
0.2076	0.2081	0.2114	0.2170	**0.2271**

Table 4: This table lists the average similarities between generated scene graphs and the ground truth. All methods are named after their visual relationship detectors.

Compare architectural choices. This study is to compare the effect of different choices in the DR-Net architecture. The choices we study here include: the number of inference units and whether the relational weights are shared across these units. The comparison is conducted on *sVG*.

Figure 5 shows the resultant curves. From the results we can see: (1) On both settings, the recall increases as the number of inference units increases. The best model can improve the recall from 56% to 73%, as the number of inference units increases. With weight sharing, the recall saturates with 12 inference units; while without sharing, the recall increases more rapidly, and saturates when it has 8 inference units. (2) Generally, with same number of inference units, the network without weight sharing performs relatively better, due to the greater expressive power.

5.3. Scene Graph Generation

Our model for visual relationship detection can be used for scene graph generation, which can serve as the basis for many tasks, *e.g.* image captioning [61, 62], visual question answering [63] and image retrieval [9].

The task here is to generate a directed graph for each image that captures objects, object attributes, and the relationships between them [9]. See Figure 6 for an illustration. We compared several configs of our method, including A_1, S, A_1S and A_1SD, with *VR* [1] on this task, on a dataset *sVG-a*, which extends *sVG* with attribute annotations. All methods are augmented with an attribute recognizer.

For each test image, we measure the similarity [64] between the generated scene graph and the ground truth. We report average similarity over all test images as our metric. Table 4 compares the results of these approaches, where A_1SD achieves the best result. This comparison indicates that with better relationship detection, one can obtain better scene graphs.

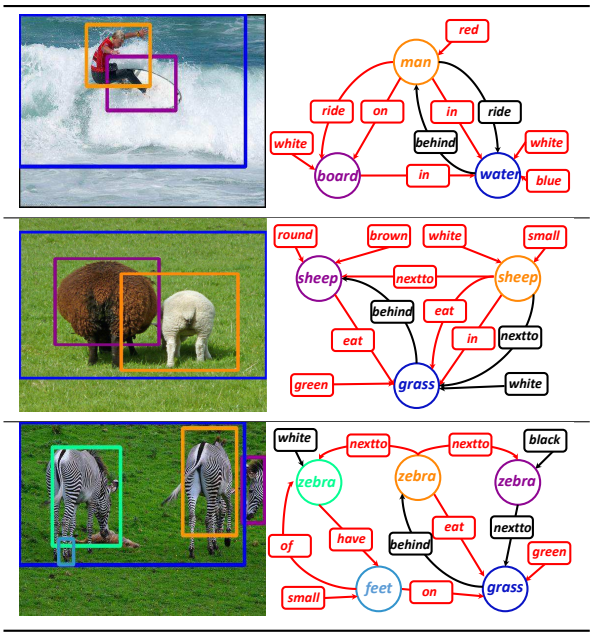


Figure 6: This figure illustrates some images and their corresponding scene graphs. The scene graphs are generated according to section 5.3. In the scene graphs, the **black** edges indicate wrong prediction, and the red edges indicate correct prediction.

6. Conclusion

This paper presented a new framework for visual relationship detection, which integrates a variety of cues: appearance, spatial configurations, as well as the statistical relations between objects and relationship predicates. At the heart of this framework is the *Deep Relational Network (DR-Net)*, a novel formulation that extends the expressive power of deep neural networks to relational modeling. On Visual Genome, the proposed method not only outperforms the state of the art by a remarkable margin, but also yields promising results in scene graph generation, a task that represents higher level of image understanding. These experimental results clearly demonstrate the significance of statistical relations in visual relationship detection, and DR-Net's strong capability in modeling complex relations.

Acknowledgment This work is partially supported by the Early Career Scheme (ECS) grant (No. 24204215), and the Big Data Collaboration grant from SenseTime Group.

References

[1] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. *arXiv preprint arXiv:1608.00187*, 2016. 1, 2, 5, 6, 7, 8

[2] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 1, 3

[3] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*, 2016. 1

[4] Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1644, 2014. 1

[5] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 1, 5

[6] 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, 3, 6

[7] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001. 2, 4

[8] Abhinav Gupta and Larry S Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *European conference on computer vision*, pages 16–29. Springer, 2008. 2

[9] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A Shamma, Michael S Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678. IEEE, 2015. 2, 4, 8

[10] Carolina Galleguillos, Andrew Rabinovich, and Serge Belongie. Object categorization using co-occurrence, location and appearance. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 2

[11] Wongun Choi, Yu-Wei Chao, Caroline Pantofaru, and Silvio Savarese. Understanding indoor scenes using 3d geometric phrases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 33–40, 2013. 2

[12] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*. Citeseer, 2011. 2

[13] Bangpeng Yao and Li Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, page 916. IEEE, 2010. 2

[14] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with r* cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1080–1088, 2015. 2

[15] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 2

[16] Jesse Thomason, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Raymond J Mooney. Integrating language and vision to generate natural language descriptions of videos in the wild. In *COLING*, volume 2, page 9, 2014. 2

[17] Vignesh Ramanathan, Congcong Li, Jia Deng, Wei Han, Zhen Li, Kunlong Gu, Yang Song, Samy Bengio, Chuck Rossenberg, and Li Fei-Fei. Learning semantic relationships for better action retrieval in images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1100–1109. IEEE, 2015. 2

[18] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. Translating video content to natural language descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 433–440, 2013. 2

[19] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2712–2719, 2013. 2

[20] Stanislaw Antol, C Lawrence Zitnick, and Devi Parikh. Zero-shot learning via visual abstraction. In *European Conference on Computer Vision*, pages 401–416. Springer, 2014. 2

[21] Mohamed Elhoseiny, Scott Cohen, Walter Chang, Brian Price, and Ahmed Elgammal. Sherlock: Scalable fact learning in images. *arXiv preprint arXiv:1511.04891*, 2015. 2

[22] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *European Conference on Computer Vision*, pages 15–29. Springer, 2010. 2

[23] Yuanjun Xiong, Kai Zhu, Dahua Lin, and Xiaoou Tang. Recognize complex events from static images by fusing deep channels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1600–1609, 2015. 2

[24] Carolina Galleguillos and Serge Belongie. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6):712–722, 2010. 2

[25] Josef Sivic, Bryan C Russell, Alexei A Efros, Andrew Zisserman, and William T Freeman. Discovering objects and their location in images. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 370–377. IEEE, 2005. 2

[26] M Pawan Kumar and Daphne Koller. Efficiently selecting regions for scene understanding. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3217–3224. IEEE, 2010. 2

[27] Myung Jin Choi, Joseph J Lim, Antonio Torralba, and Alan S Willsky. Exploiting hierarchical context on a large database of object categories. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 129–136. IEEE, 2010. 2

[28] Lubor Ladicky, Chris Russell, Pushmeet Kohli, and Philip HS Torr. Graph cut based inference with co-occurrence statistics. In *European Conference on Computer Vision*, pages 239–253. Springer, 2010. 2

[29] Ruslan Salakhutdinov, Antonio Torralba, and Josh Tenenbaum. Learning to share visual appearance for multiclass object detection. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1481–1488. IEEE, 2011. 2

[30] Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge Belongie. Objects in context. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 2

[31] Sanja Fidler and Ales Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 2

[32] 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 *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1605–1614. IEEE, 2006. 2

[33] Thomas Mensink, Efstratios Gavves, and Cees GM Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2441–2448, 2014. 2

[34] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision*, 106(2):210233, 2014. 2

[35] C Lawrence Zitnick, Devi Parikh, and Lucy Vanderwende. Learning the visual interpretation of sentences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1681–1688, 2013. 2

[36] Derek Hoiem, Alexei A Efros, and Martial Hebert. Putting objects in perspective. *International Journal of Computer Vision*, 80(1):3–15, 2008. 2

[37] Angel X Chang, Manolis Savva, and Christopher D Manning. Semantic parsing for text to 3d scene generation. *ACL 2014*, page 17, 2014. 2

[38] Jian Yao, Sanja Fidler, and Raquel Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 702–709. IEEE, 2012. 2

[39] Hamid Izadinia, Fereshteh Sadeghi, and Ali Farhadi. Incorporating scene context and object layout into appearance modeling. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 232–239. IEEE, 2014. 2

[40] Stephen Gould, Jim Rodgers, David Cohen, Gal Elidan, and Daphne Koller. Multi-class segmentation with relative location prior. *International Journal of Computer Vision*, 80(3):300–316, 2008. 2

[41] Alexander C Berg, Tamara L Berg, Hal Daume, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Aneesh Sood, Karl Stratos, et al. Understanding and predicting importance in images. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3562–3569. IEEE, 2012. 2

[42] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2641–2649, 2015. 2

[43] Andrej Karpathy, Armand Joulin, and Fei Fei F Li. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897, 2014. 2

[44] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. *arXiv preprint arXiv:1511.03745*, 2015. 2

[45] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2634–2641, 2013. 2

[46] Santosh K Divvala, Ali Farhadi, and Carlos Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3270–3277, 2014. 2

[47] Ramakrishna Vedantam, Xiao Lin, Tanmay Batra, C Lawrence Zitnick, and Devi Parikh. Learning common sense through visual abstraction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2542–2550, 2015. 2

[48] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1473–1482, 2015. 2, 6

[49] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 3, 4, 6, 7

[50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 3, 4, 7

[51] 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. 4, 5

[52] Ariadna Quattoni, Michael Collins, and Trevor Darrell. Conditional random fields for object recognition. In *Advances in neural information processing systems*, pages 1097–1104, 2004. 4

[53] Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Adv. Neural Inf. Process. Syst*, 2011. 4

[54] Judea Pearl. Probabilistic reasoning in intelligent systems: Networks of plausible reasoning, 1988. 4

[55] Liang-Chieh Chen, Alexander G Schwing, Alan L Yuille, and Raquel Urtasun. Learning deep structured models. In *Proc. ICML*, 2015. 5

[56] Alexander G Schwing and Raquel Urtasun. Fully connected deep structured networks. *arXiv preprint arXiv:1503.02351*, 2015. 5

[57] David Belanger and Andrew McCallum. Structured prediction energy networks. *arXiv preprint arXiv:1511.06350*, 2015. 5

[58] Zhirong Wu, Dahua Lin, and Xiaoou Tang. Deep markov random field for image modeling. In *European Conference on Computer Vision*, pages 295–312. Springer, 2016. 5

[59] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 6

[60] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 6

[61] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. *arXiv preprint arXiv:1607.08822*, 2016. 8

[62] Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. From images to sentences through scene description graphs using commonsense reasoning and knowledge. *arXiv preprint arXiv:1511.03292*, 2015. 8

[63] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Visual question answering: A survey of methods and datasets. *arXiv preprint arXiv:1607.05910*, 2016. 8

[64] Pierre-Antoine Champin and Christine Solnon. Measuring the similarity of labeled graphs. In *International Conference on Case-Based Reasoning*, pages 80–95. Springer, 2003. 8