Visual Relationship Detection with Language Priors

Cewu Lu*, Ranjay Krishna*, Michael Bernstein, Li Fei-Fei
{cwlu, ranjaykrishna, msb, feifeili}@cs.stanford.edu

Stanford University

Abstract. Visual relationships capture a wide variety of interactions between pairs of objects in images (e.g. "man riding bicycle" and "man pushing bicycle"). Consequently, the set of possible relationships is extremely large and it is difficult to obtain sufficient training examples for all possible relationships. Because of this limitation, previous work on visual relationship detection has concentrated on predicting only a handful of relationships. Though most relationships are infrequent, their objects (e.g. "man" and "bicycle") and predicates (e.g. "riding" and "pushing") independently occur more frequently. We propose a model that uses this insight to train visual models for objects and predicates individually and later combines them together to predict multiple relationships per image. We improve on prior work by leveraging language priors from semantic word embeddings to finetune the likelihood of a predicted relationship. Our model can scale to predict thousands of types of relationships from a few examples. Additionally, we localize the objects in the predicted relationships as bounding boxes in the image. We further demonstrate that understanding relationships can improve content based image retrieval.

1 Introduction

While objects are the core building blocks of an image, it is often the relationships between objects that determine the holistic interpretation. For example, an image with a `person` and a `bicycle` might involve the man `riding`, `pushing`, or even `falling off` of the bicycle (Figure 1). Understanding this diversity of relationships is central to accurate image retrieval and to a richer semantic understanding of our visual world.

Visual relationships are a pair of localized objects connected via a predicate (Figure 2). We represent relationships as $\langle \texttt{object}_1 \text{ - } \texttt{predicate} \text{ - } \texttt{object}_2 \rangle$ [1]. Visual relationship detection involves detecting and localizing pairs of objects in an image and also classifying the predicate or interaction between each pair (Figure 2). While it poses similar challenges as object detection [1], one critical difference is that the size of the semantic space of possible relationships is much larger than that of objects. Since relationships are composed of two objects, there is a greater skew of rare relationships as object co-occurrence is infrequent in

* = equal contribution

Fig. 1: Even though all the images contain the same objects (a `person` and a `bicycle`), it is the relationship between the objects that determine the holistic interpretation of the image.

images. So, a fundamental challenge in visual relationship detection is learning from very few examples.

Visual Phrases [6] studied visual relationship detection using a small set of 13 common relationships. Their model requires enough training examples for every possible $\langle$`object`$_1$ - `predicate` - `object`$_2\rangle$ combination, which is difficult to collect owing to the infrequency of relationships. If we have N objects and K predicates, Visual Phrases [6] would need to train $\mathcal{O}(N^2K)$ unique detectors separately. We use the insight that while relationships (e.g. "person jumping over a fire hydrant") might occur rarely in images, its objects (e.g. `person` and `fire hydrant`) and predicate (e.g. `jumping over`) independently appear more frequently. We propose a **visual appearance module** that learns the appearance of objects and predicates and fuses them together to jointly predict relationships. We show that our model only needs $\mathcal{O}(N+K)$ detectors to detect $\mathcal{O}(N^2K)$ relationships.

Another key observation is that relationships are semantically related to each other. For example, a "person riding a horse" and a "person riding an elephant" are semantically similar because both `elephant` and `horse` are animals. Even if we haven't seen many examples of "person riding an elephant", we might be able to infer it from a "person riding a horse". Word vector embeddings [7] naturally lend themselves in linking such relationships because they capture semantic similarity in language (e.g. `elephant` and `horse` are cast close together in a word vector space). Therefore, we also propose a **language module** that uses pre-trained word vectors [7] to cast relationships into a vector space where similar relationships are optimized to be close to each other. Using this embedding space, we can finetune the prediction scores of our relationships and even enable zero shot relationship detection.

In this paper, we propose a model that can learn to detect visual relationships by (1) (1) learning visual appearance models for its objects and predicates and (2) using the relationship embedding space learnt from language. We train our model by optimizing a bi-convex function. To benchmark the task of visual relationship detection, we introduce a new dataset that contains 5000 images with 37,993 relationships. Existing datasets that contain relationships were designed

[1] In natural language processing [2,3,4,5], relationships are defined as $\langle$`subject - predicate - object`$\rangle$. In this paper, we define them as $\langle$`object`$_1$ - `predicate - object`$_2\rangle$ for simplicity.

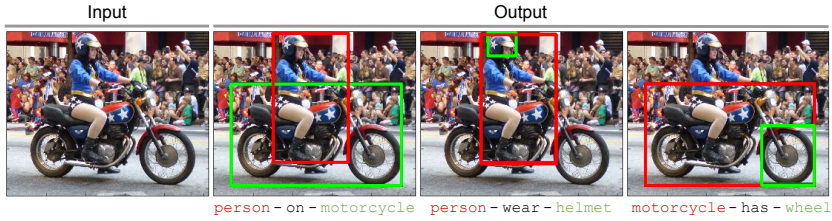

Fig. 2: Visual Relationship Detection: Given an image as input, we detect multiple relationships in the form of ⟨object$_1$ - relationship - object$_2$⟩. Both the objects are localized in the image as bounding boxes. In this example, we detect the following relationships: ⟨person - on - motorcycle⟩, ⟨person - wear - helmet⟩ and ⟨motorcycle - has - wheel⟩.

for improving object detection [6] or image retrieval [8] and hence, don't contain sufficient variety of relationships or predicate diversity per object category. Our model outperforms all previous models in visual relationship detection. We further study how our model can be used to perform zero shot visual relationship detection. Finally, we demonstrate that understanding relationships can improve image-based retrieval.

2 Related Work

Visual relationship prediction involves detecting the objects that occur in an image as well as understanding the interactions between them. There has been a series of work related to improving object detection by leveraging **object co-occurrence** statistics [9,10,11,12,13,14]. Structured learning approaches have improved scene classification along with object detection using hierarchial contextual data from co-occurring objects [15,16,17,18]. Unlike these methods, we study the *context* or *relationships* in which these objects co-occur.

Some previous work has attempted to learn **spatial relationships** between objects [19,13] to improve segmentation [19]. They attempted to learn four spatial relationships: "above", "below", "inside", and "around" [13]. While we believe that that learning spatial relationships is important, we also study non-spatial relationships such as pull (actions), taller than (comparative), etc.

There have been numerous efforts in **human-object interaction** [20,21,22] and action recognition [23] to learn discriminative models that distinguish between relationships where object$_1$ is a human (e.g. "playing violin" [24]). Visual relationship prediction is more general as object$_1$ is not constrained to be a human and the predicate doesn't have to be a verb.

Visual relationships are not a new concept. Some papers explicitly collected relationships in images [25,26,27,28,29] and videos [27,30,31] and helped models map these relationships from images to language. Relationships have also improved object localization [32,33,6,34]. A meaning space of relationships have aided the cognitive task of mapping images to captions [35,36,37,38]. Finally, they have been used to generate indoor images from sentences [39] and

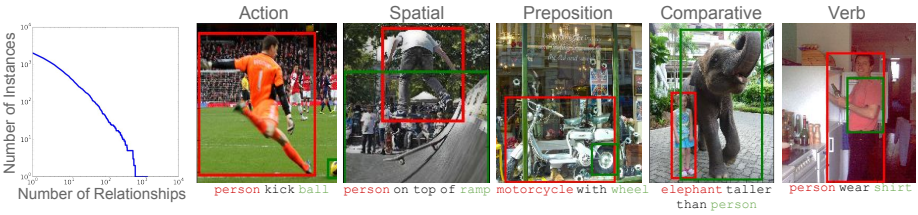

Fig. 3: (left) A log scale distribution of the number of instances to the number of relationships in our dataset. Only a few relationships occur frequently and there is a long tail of infrequent relationships. (right) Relationships in our dataset can be divided into many categories, 5 of which are shown here: verb, spatial, preposition, comparative and action.

Table 1: Comparison between our visual relationship benchmarking dataset with existing datasets that contain relationships. Relationships and Objects are abbreviated to Rel. and Obj. because of space constraints.

	Images	Rel. Types	Rel. Instances	# Predicates per Obj. Category
Visual Phrases [6]	2,769	13	2,040	120
Scene Graph [8]	5,000	23,190	109,535	2.3
Ours	5,000	6,672	37,993	24.25

to improve image search [8,40]. In this paper, we formalize visual relationship prediction as a task onto itself and demonstrate further improvements in image retrieval.

The most recent attempt at relationship prediction has been in the form of **visual phrases**. Learning appearance models for visual phrases has shown to improve individual object detection, i.e. detecting "a person riding a horse" improves the detection and localization of "person" and "horse" [6,41]. Unlike our model, all previous work has attempted to detect only a handful of visual relationships and do not scale because most relationships are infrequent. We propose a model that manages to scale and detect millions of types of relationships. Additionally, our model is able to detect unseen relationships.

3 Visual Relationship Dataset

Visual relationships put objects in context; they capture the different interactions between pairs of objects. These interactions (shown in Figure 3) might be verbs (e.g. `wear`), spatial (e.g. `on top of`), prepositions (e.g. `with`), comparative (e.g. `taller than`), actions (e.g. `kick`) or a preposition phrase (e.g. `drive on`). A dataset for visual relationship prediction is fundamentally different from a dataset for object detection. A relationship dataset should contain more than just objects localized in images; it should capture the rich variety of interactions between pairs of objects (predicates per object category). For example, a `person`

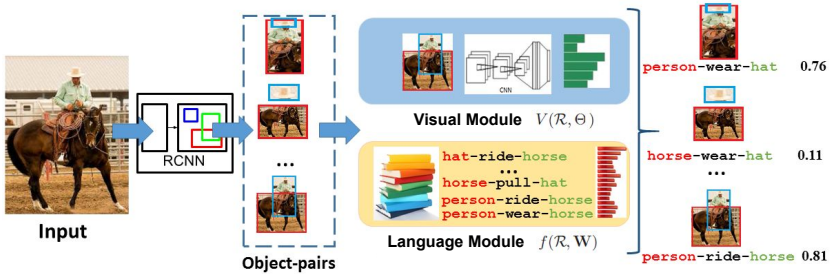

Fig. 4: A overview of our visual relationship detection pipeline. Given an image as input, RCNN [43] generates a set of object proposals. Each pair of object proposals is then scored using a (1) visual appearance module and a (2) language module. These scores are then thresholded to output a set of relationship labels (e.g. ⟨`person - riding - horse`⟩). Both objects in a relationship (e.g. `person` and `horse`) are localized as bounding boxes. The parameters of those two modules (W and Θ) are iteratively learnt in Section 4.1.

can be associated with predicates such as `ride`, `wear`, `kick` etc. Additionally, the dataset should contain a large number of possible relationships types.

Existing datasets that contain relationships were designed to improve object detection [6] or image retrieval [8]. The Visual Phrases [6] dataset focuses on 17 common relationship types. But, our goal is to understand the rich variety of infrequent relationships. On the other hand, even though the Scene Graph dataset [8] has 23,190 relationship types [2], it only has 2.3 predicates per object category. Detecting relationships on the Scene Graph dataset [8] essentially boils down to object detection. Therefore, we designed a dataset specifically for benchmarking visual relationship prediction.

Our dataset (Table 1) contains 5000 images with 100 object categories and 70 predicates. In total, the dataset contains 37,993 relationships with 6,672 relationship types and 24.25 predicates per object category. Some example relationships are shown in Figure 3. The distribution of relationships in our dataset highlights the long tail of infrequent relationships (Figure 3(left)). We use 4000 images in our training set and test on the remaining 1000 images. 1,877 relationships occur in the test set but never occur in the training set.

4 Visual Relationship Prediction Model

The goal of our model is to detect visual relationships from an image. During training (Section 4.1), the input to our model is a fully supervised set of images

[2] Note that the Scene Graph dataset [8] was collected using unconstrained language, resulting in multiple annotations for the same relationship (e.g. ⟨`man - kick - ball`⟩ and ⟨`person - is kicking - soccer ball`⟩). Therefore, 23,190 is an inaccurate estimate of the number of unique relationship types in their dataset. We do not compare with the Visual Genome dataset [42] because their relationships had not been released at the time this paper was written.

with relationship annotations where the objects are localized as bounding boxes and labelled as $\langle \texttt{object}_1 \text{ - } \texttt{predicate} \text{ - } \texttt{object}_2 \rangle$. At test time (Section 4.2), our input is an image with no annotations. We predict multiple relationships and localize the objects in the image. Figure 4 illustrates a high level overview of our detection pipeline.

4.1 Training Approach

In this section, we describe how we train our visual appearance and language modules. Both the modules are combined together in our objective function.

Visual Appearance Module While Visual Phrases [6] learned a separate detector for every single relationship, we model the appearance of visual relationships $V()$ by learning the individual appearances of its comprising objects and predicate. While relationships are infrequent in real world images, the objects and predicates can be learnt as they independently occur more frequently. Furthermore, we demonstrate that our model outperforms Visual Phrases' detectors, showing that learning individual detectors outperforms learning detectors for relationships together (Table 2).

First, we train a convolutional neural network (CNN) (VGG net [44]) to classify each of our $N = 100$ objects. Similarly, we train a second CNN (VGG net [44]) to classify each of our $K = 70$ predicates using the union of the bounding boxes of the two participating objects in that relationship. Now, for each ground truth relationship $R_{\langle i,k,j \rangle}$ where i and j are the object classes (with bounding boxes O_1 and O_2) and k is the predicate class, we model V (Figure 4) as:

$$V(R_{\langle i,k,j \rangle}, \Theta | \langle O_1, O_2 \rangle) = P_i(O_1)(\mathbf{z}_k^T \text{CNN}(O_1, O_2) + s_k) P_j(O_2) \quad (1)$$

where Θ is the parameter set of $\{\mathbf{z}_k, s_k\}$. $\mathbf{z}_k$ and s_k are the parameters learnt to convert our CNN features to relationship likelihoods. $k = 1, \ldots, K$ represent the K predicates in our dataset. $P_i(O_1)$ and $P_j(O_2)$ are the CNN likelihoods of categorizing box O_1 as object category i and box O_2 as category j. $\text{CNN}(O_1, O_2)$ is the predicate CNN features extracted from the union of the O_1 and O_2 boxes.

Language Module One of our key observations is that relationships are semantically related to one another. For example, $\langle \texttt{person - ride - horse} \rangle$ is semantically similar to $\langle \texttt{person - ride - elephant} \rangle$. Even if we have not seen any examples of $\langle \texttt{person - ride - elephant} \rangle$, we should be able to infer it from similar relationships that occur more frequently (e.g. $\langle \texttt{person - ride - horse} \rangle$). Our language module projects relationships into an embedding space where similar relationships are optimized to be close together. We first describe the function that projects a relationship to the vector space (Equation 2) and then explain how we train this function by enforcing similar relationships to be close together in a vector space (Equation 4) and by learning a likelihood prior on relationships (Equation 5).

Projection Function First, we use pre-trained word vectors (word2vec) [7] to cast the two objects in a relationship into an word embedding space [7]. Next, we concatenate these two vectors together and transform it into the relationship vector space using a projection parameterized by $\mathbf{W}$, which we learn. This projection presents how two objects interact with each other. We denote *word2vec*() as the function that converts a word to its 300 *dim.* vector. The relationship projection function (shown in Figure 4) is defined as:

$$f(\mathcal{R}_{\langle i,k,j\rangle}, \mathbf{W}) = \mathbf{w}_k^T[word2vec(t_i), word2vec(t_j)] + b_k \tag{2}$$

where t_j is the word (in text) of the j^{th} object category. $\mathbf{w}_k$ is a 600 *dim.* vector and b_k is a bias term. $\mathbf{W}$ is the set of $\{\{\mathbf{w}_1, b_1\}, \ldots, \{\mathbf{w}_k, b_k\}\}$, where each row presents one of our K predicates.

Training Projection Function We want to optimize the projection function $f()$ such that it projects similar relationships closer to one another. For example, we want the distance between ⟨`man - riding - horse`⟩ to be close to ⟨`man - riding - cow`⟩ but farther from ⟨`car - has - wheel`⟩. We formulate this by using a heuristic where the distance between two relationships is proportional to the word2vec distance between its component objects and predicate:

$$\frac{[f(\mathcal{R}, \mathbf{W}) - f(\mathcal{R}', \mathbf{W})]^2}{d(\mathcal{R}, \mathcal{R}')} = constant, \quad \forall \mathcal{R}, \mathcal{R}' \tag{3}$$

where $d(\mathcal{R}, \mathcal{R}')$ is the sum of the cosine distances (in word2vec space [7]) between of the two objects and the predicates of the two relationships $\mathcal{R}$ and $\mathcal{R}'$. Now, to satisfy Eq 3, we randomly sample pairs of relationships ($\langle \mathcal{R}, \mathcal{R}' \rangle$) and minimize their variance:

$$K(\mathbf{W}) = var(\{\frac{[f(\mathcal{R}, \mathbf{W}) - f(\mathcal{R}', \mathbf{W})]^2}{d(\mathcal{R}, \mathcal{R}')} \quad \forall \mathcal{R}, \mathcal{R}'\}) \tag{4}$$

where $var()$ is a variance function. The sample number we use is 500K.

Likelihood of a Relationship The output of our projection function should ideally indicate the likelihood of a visual relationship. For example, our model should not assign a high likelihood score to a relationship like ⟨`dog - drive - car`⟩, which is unlikely to occur. We model this by enforcing that if $\mathcal{R}$ occurs more frequently than $\mathcal{R}'$ in our training data, then it should have a higher likelihood of occurring again. We formulate this as a rank loss function:

$$L(\mathbf{W}) = \sum_{\{\mathcal{R}, \mathcal{R}'\}} \max\{f(\mathcal{R}', \mathbf{W}) - f(\mathcal{R}, \mathbf{W}) + 1, 0\} \tag{5}$$

While we only enforce this likelihood prior for the relationships that occur in our training data, the projection function $f()$ generalizes it for all ⟨`object`$_1$ `- predicate - object`$_2$⟩ combinations, even if they are not present in our training data. The max operator here is to encourage correct ranking (with margin) $f(\mathcal{R}, \mathbf{W}) - f(\mathcal{R}', \mathbf{W}) \geq 1$. Minimizing this objective enforces that a relationship with a lower likelihood of occurring has a lower $f()$ score.

Objective function So far we have presented our visual appearance module ($V()$) and the language module ($f()$). We combine them to maximize the rank of the ground truth relationship $\mathcal{R}$ with bounding boxes O_1 and O_2 using the following rank loss function:

$$C(\Theta, \mathbf{W}) = \sum_{\langle O_1 O_2\rangle, \mathcal{R}} \max\{1 - V(\mathcal{R}, \Theta|\langle O_1, O_2\rangle) f(\mathcal{R}, \mathbf{W}) + \max_{\langle O_1', O_2'\rangle \neq \langle O_1, O_2\rangle, \mathcal{R}' \neq \mathcal{R}} V(\mathcal{R}', \Theta|\langle O_1', O_2'\rangle) f(\mathcal{R}', \mathbf{W}), 0\} \quad (6)$$

We use a ranking loss function to make it more likely for our model to choose the correct relationship. Given the large number of possible relationships, we find that a classification loss performs worse. Therefore, our final objective function combines Eq 6 with Eqs 4 and 5 as:

$$\min_{\Theta, \mathbf{W}} \{C(\Theta, \mathbf{W}) + \lambda_1 L(\mathbf{W}) + \lambda_2 K(\mathbf{W})\} \quad (7)$$

where $\lambda_1 = 0.05$ and $\lambda_2 = 0.002$ are hyper-parameters that were obtained though grid search to maximize performance on the validation set. Note that both Eqs 6 and 5 are convex functions. Eq 4 is a biqudratic function with respect to $\mathbf{W}$. So our objective function Eq 7 has a quadratic closed form. We perform stochastic gradient descent iteratively on Eqs 6 and 5. It converges in $20 \sim 25$ iterations.

4.2 Testing

At test time, we use RCNN [43] to produce a set of candidate object proposals for every test image. Next, we use the parameters learnt from the visual appearance model (Θ) and the language module ($\mathbf{W}$) to predict visual relationships ($\mathcal{R}^*_{\langle i,k,j\rangle}$) for every pair of RCNN object proposals $\langle O_1, O_2\rangle$ using:

$$\mathcal{R}^* = \arg\max_{\mathcal{R}} V(\mathcal{R}, \Theta|\langle O_1, O_2\rangle) f(\mathcal{R}, \mathbf{W}) \quad (8)$$

5 Experiments

We evaluate our model by detecting visual relationships from images. We show that our proposed method outperforms previous state-of-the-art methods on our dataset (Section 5.1) as well as on previous datasets (Section 5.3). We also measure how our model performs in zero-shot learning of visual relationships (Section 5.2). Finally, we demonstrate that understanding visual relationship can improve common computer vision tasks like content based image retrieval (Section 5.4).

5.1 Visual Relationship Detection

Setup. Given an input image, our task is to extract a set of visual relationships $\langle$`object`$_1$ - `predicate` - `object`$_2\rangle$ and localize the objects as bounding boxes

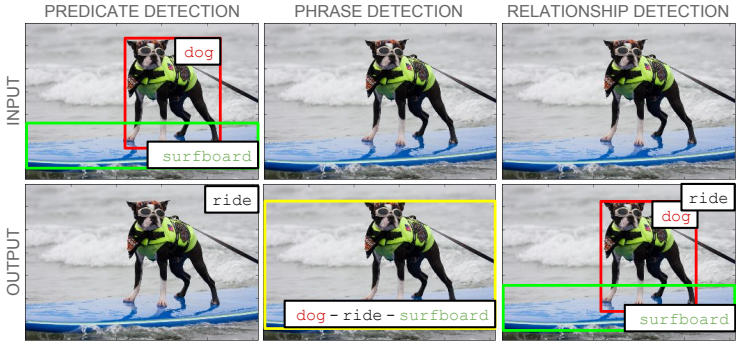

Fig. 5: We evaluate visual relationship detection using three conditions: predicate detection (where we only predict the predicate given the object classes and boxes), phrase detection (where we label a region of an image with a relationship) and relationship detection (where we detect the objects and label the predicate between them).

in the image. We train our model using the 4000 training images and perform visual relationship prediction on the 1000 test images.

The evaluation metrics we report is **recall @ 100** and **recall @ 50** [45]. **Recall @ x** computes the fraction of times the correct relationship is predicted in the top **x** confident relationship predictions. Since we have 70 predicates and an average of 18 objects per image, the total possible number of relationship predictions is $100 \times 70 \times 100$, which implies that the random guess will result in a recall @ 100 of 0.00014. We notice that mean average precision (mAP) is another widely used metric. However, mAP is a pessimistic evaluation metric because we can not exhaustively annotate all possible relationships in an image. Consider the case where our model predicts $\langle$`person - taller than - person`$\rangle$. Even if the prediction is correct, mAP would penalize the prediction if we do not have that particular ground truth annotation.

Detecting a visual relationship involves classifying both the objects, predicting the predicate and localization both the objects. To study how our model performs on each of these tasks, we measure visual relationship prediction under the following conditions:

1. In **predicate detection** (Figure 5(left)), our input is an image and set of localized objects. The task is to predict a set of possible predicates between pairs of objects. This condition allows us to study how difficult it is to predict relationships without the limitations of object detection [43].
2. In **phrase detection** (Figure 5(middle)), our input is an image and our task is to output a label $\langle$`object`$_1$ `- predicate -` `object`$_2\rangle$ and localize the entire relationship as *one* bounding box having at least 0.5 overlap with ground truth box. This is the evaluation used in Visual Phrases [6].
3. In **relationship detection** (Figure 5(right)), our input is an image and our task is to output a set of $\langle$`object`$_1$ `- predicate -` `object`$_2\rangle$ and localize *both* object$_1$ and object$_2$ in the image having at least 0.5 overlap with their ground truth boxes simultaneously.

Comparison Models. We compare our method with some state-of-that-art approaches [6,44]. We further perform ablation studies on our model, considering just the visual appearance and the language module, including the likelihood term (Eq 4) and embedding term (Eq 5) to study their contributions.

- **Visual phrases.** Similar to Visual Phrases [6], we train deformable parts models for each of the 6,672 relationships (e.g. "chair under table") in our training set.
- **Joint CNN.** We train a CNN model [44] to predict the three components of a relationship together. Specifically, we train a 270 $(100 + 100 + 70)$ way classification model that learns to score the two objects (100 categories each) and predicate (70 categories). This model represents the Visual phrases
- **Visual appearance (Ours - V only).** We only use the visual appearance module of our model described in Eq 6 by optimizing $V()$.
- **Likelihood of a relationship (Ours - L only).** We only use the likelihood of a relationship described in Eq 5 by optimizing $L()$.
- **Visual appearance + naive frequency (Ours - V + naive FC).** One of the contributions of our model is the ability to use a language prior via our semantic projection function $f()$ (Eq 2). Here, we replace $f()$ with a function that maps a relationship to its frequency in our training data. Using this naive function, we hope to test the effectiveness of $f()$.
- **Visual appearance + Likelihood (Ours - V + L only).** We use both the visual appearance module (Eq 6) and the likelihood term (Eq 5) by optimizing both $V()$ and $L()$. The only part of our model missing is $K()$ Eq 4, which projects similar relationships closer.
- **Visual appearance + likelihood + regularizer (Ours - V + L + Reg.).** We use the visual appearance module and the likelihood term and add an L_2 regularizer on W.
- **Full Model (Ours - V + L + K).** This is our full model. It contains the visual appearance module (Eq 6), the likelihood term (Eq 5) and the embedding term (Eq 4) from similar relationships.

Results. Visual Phrases [6] and Joint CNN [44] train an individual detector for every relationship. Since the space of all possible relationships is large (we have 6,672 relationship types in the training set), there is a shortage of training examples for infrequent relationships, causing both models to perform poorly on predicate, phrase and relationship detection (Table 2). (Ours - V only) can't discriminative between similar relationships by itself resulting in 1.85 R@100 for relationship detection. Similarly, (Ours - L only) always predicts the most frequent relationship ⟨`person - wear - shirt`⟩ and results in 0.08 R@100, which is the percentage of the most frequent relationship in our testing data. These problems are remedied when both V and L are combined in (Ours - V + L only) with an increase of 3% R@100 in on both phrase and relationship detection and more than 10% increase in predicate detection. (V + Naive FC) is missing our relationship projection function $f()$, which learns the likelihood of a predicted relationship and performs worse than (Ours - V + L only) and (Ours -

Table 2: Results for visual relationship detection (Section 5.1). R@100 and R@50 are abbreviations of Recall @ 100 and Recall @ 50. Note that in predicate det., we are predicting multiple predicates per image (one between every pair of objects) and hence R@100 is less than 1.

	Phrase Det.		Relationship Det.		Predicate Det.	
	R@100	R@50	R@100	R@50	R@100	R@50
Visual Phrases [6]	0.07	0.04	-	-	1.91	0.97
Joint CNN [44]	0.09	0.07	0.09	0.07	2.03	1.47
Ours - V only	2.61	2.24	1.85	1.58	7.11	7.11
Ours - L only	0.08	0.08	0.08	0.08	18.22	18.22
Ours - V + naive FC	6.39	6.65	5.47	5.27	28.87	28.87
Ours - V + L only	8.59	9.13	9.18	9.04	35.20	35.20
Ours - V + L + Reg.	8.91	9.60	9.63	9.71	36.31	36.31
Ours - V + L + K	**17.03**	**16.17**	**14.70**	**13.86**	**47.87**	**47.87**

V + L + K). Also, we observe that (Ours - V + L + K) has an 11% improvement in comparison to (Ours - V + L only) in predicate detection, demonstrating that the language module from similar relationships significantly helps improve visual relationship detection. Finally, (Ours - V + L + K) outperforms (Ours - V + L + Reg.) showcasing the $K()$ is acting not only as a regularizer but is learning to preserve the distances between similar relationships.

By comparing the performance of all the models between relationship and predicate detection, we notice a 30% drop in R@100. This drop in recall is largely because we have to localize two objects simultaneously, amplifying the object detection errors. Note that even when we have ground truth object proposals (in predicate detection), R@100 is still 47.87.

Qualitative Results. In Figure 6(a)(b)(c), Visual Phrase and Joint CNN incorrectly predict a common relationship: ⟨`person - drive - car`⟩ and ⟨`car - next to - tree`⟩. These models tend to predict the most common relationship as they see a lot of them during training. In comparison, our model correctly predicts and localizes the objects in the image. Figure 6(d)(e)(f) compares the various components of our model. Without the relationship likelihood score, (Ours - V only) incorrectly classifies a `wheel` as a clock in (d) and mislabels the predicate in (e) and (f). Without any visual priors, (Ours - L only) always reports the most frequent relationship ⟨`person - wear - shirt`⟩. (Ours - V + L) fixes (d) by correcting the visual model's misclassification of the `wheel` as a `clock`. But it still does not predict the correct predicate for (e) and (f) because ⟨`person - ride - elephant`⟩ and ⟨`hand - hold - phone`⟩ rarely occur in our training set. However, our full model (Ours - V + L + K) leverages similar relationships it has seen before and is able to correctly detect the relationships in (e) and (f).

5.2 Zero-shot Learning

Owing to the long tail of relationships in real world images, it is difficult to build a dataset with every possible relationship. Therefore, a model that detects visual

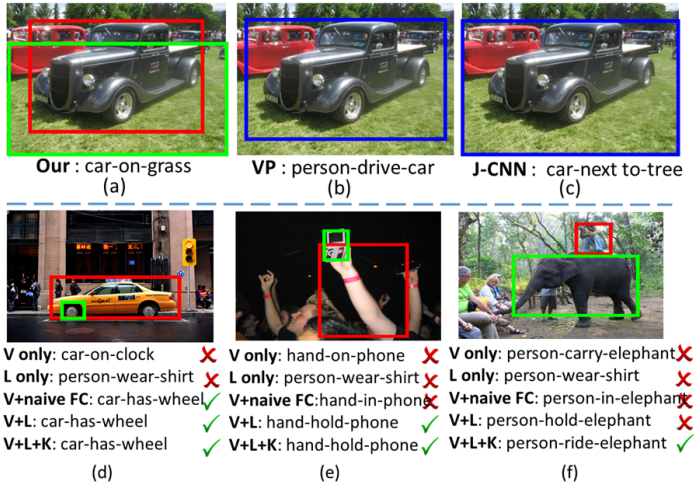

Fig. 6: (a), (b) and (c) show results from our model, Visual Phrases [6] and Joint CNN [44] on the same image. All ablation studies results for (d), (e) and (f) are reported below the corresponding image. Ticks and crosses mark the correct and incorrect results respectively. Phrase, object$_1$ and object$_2$ boxes are in blue, red and green respectively.

Table 3: Results for zero-shot visual relationship detection (Section 5.2). Visual Phrases, Joint CNN and Ours - V + naive FC are omitted from this experiment as they are unable to do zero-shot learning.

	Phrase Det.		Relationship Det.		Predicate Det.	
	R@100	R@50	R@100	R@50	R@100	R@50
Ours - V only	1.12	0.95	0.78	0.67	3.52	3.52
Ours - L only	0.01	0.00	0.01	0.00	5.09	5.09
Ours - V + L only	2.56	2.43	2.66	2.27	6.11	6.11
Ours - V + L + K	**3.75**	**3.36**	**3.52**	**3.13**	**8.45**	**8.45**

relationships should also be able to perform zero-shot prediction of relationships it has never seen before. Our model is able to leverage similar relationships it has already seen to detect unseen ones.

Setup. Our test set contains 1,877 relationships that never occur in our training set (e.g. ⟨`elephant - stand on - street`⟩). These unseen relationships can be inferred by our model using similar relationships (e.g. ⟨`dog - stand on - street`⟩) from our training set. We report our results for detecting unseen relationships in Table 3 for predicate, phrase, and relationship detection.

Results. (Ours - V) achieves a low 3.52 R@100 in predicate detection because visual appearances are not discriminative enough to predict unseen relationships. (Ours - L only) performs poorly in predicate detection (5.09 R@100) because it automatically returns the most common predicate. By comparing (Ours - V + L+ K) and (Ours - V + L only), we find the use of K gains an improvement of 30% since it utilizes similar relationships to enable zero shot predictions.

Table 4: Visual phrase detection results on Visual Phrases dataset [6].

	Phrase Detection			Zero-Shot Phrase Detection		
	R@100	R@50	mAP	R@100	R@50	mAP
Visual Phrase [6]	52.7	49.3	38.0	-	-	-
Joint CNN	75.3	71.5	54.1	-	-	-
Ours V only	72.0	68.6	53.4	13.5	11.3	5.3
Ours V + naive FC	77.8	73.4	55.8	-	-	-
Ours V + L only	79.3	76.7	57.3	17.8	15.1	8.8
Ours V + L + K	**82.7**	**78.1**	**59.2**	**11.4**	**23.9**	**18.5**

5.3 Visual Relationship Detection on Existing Dataset

Our goal in this paper is to understand the rich variety of infrequent relationships. Our comparisons in Section 3 show that existing datasets either do not have enough diveristy of predicates per object category or enough relationship types. Therefore, we introduced a new dataset (in Section 3) and tested our visual relationship detection model in Section 5.1 and Section 5.2. In this section, we run additional experiments on the existing visual phrases dataset [6] to provide further benchmarks.

Setup. The visual phrase dataset contains 17 phrases (e.g. "dog jumping"). We evaluate the models (introduced in Section 5.1) for visual relationship detection on 12 of these phrases that can be represented as a $\langle \texttt{object}_1 \text{ - } \texttt{predicate} \text{ - } \texttt{object}_2 \rangle$ relationship. To study zero-shot learning, we remove two phrases ("person lying on sofa" and "person lying on beach") from the training set, and attempt to recognize them in the testing set. We report mAP, R@50 and R@100.

Results. In Table 4 we see that our method is able to perform better than the existing Visual Phrases' model even though the dataset is small and contains only 12 relationships. We get a mAP of 0.59 using our entire model as compared to a mAP of 0.38 using Visual Phrases' model. We also outperform the Joint CNN baseline, which achieves a mAP of 0.54. Considering that (Ours - V only) model performs similarly to the baselines, we believe that our full model's improvements on this dataset are heavily influenced by the language priors. By learning to embed similar relationships close to each other, the language model's aid can be thought of as being synonymous to the improvements achieved through training set augmentation. Finally, we see a similar improvements in zero shot learning.

5.4 Image based Retrieval

An important task in computer vision is image retrieval. An improved retrieval model should be able to infer the relationships between objects in images. We will demonstrate that the use of visual relationships can improve retrieval quality.

Setup. Recall that our test set contains 1000 images. Every query uses 1 of these 1000 images and ranks the remaining 999. We use 54 query images in our experiments. Two annotators were asked to rank image results for each of the 54 queries. To avoid bias, we consider the results for a particular query as ground

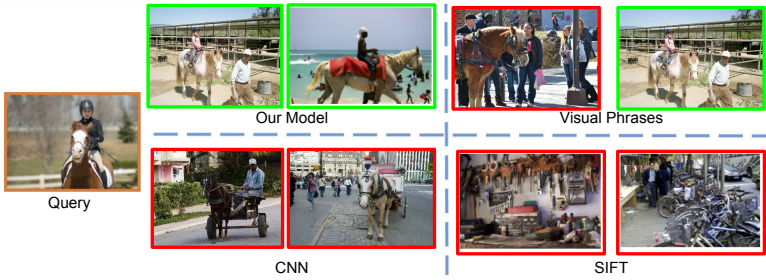

Fig. 7: Examples retrieval results using an image as the query.

Table 5: Example image retrieval using a image of a ⟨person - ride - horse⟩ (Section 5.4). Note that a *higher* recall and *lower* median rank indicates better performance.

	Recall @ 1	Recall @ 5	Recall @ 10	Median Rank
GIST [46]	0.00	5.60	8.70	68
SIFT [47]	0.70	6.10	10.3	54
CNN [44]	3.15	7.70	11.5	20
Visual Phrases [6]	8.72	18.12	28.04	12
Our Model	**10.82**	**30.02**	**47.00**	**4**

truth only if it was selected by both annotators. We evaluate performance using R@1, R@5 and R@10 and median rank [8]. For comparison, we use three image descriptors that are commonly used in image retrieval: CNN [44], GIST [46] and SIFT [47]. We rank results for a query using the L_2 distance from the query image. Given a query image, our model predicts a set of visual relationships $\{R_1, \ldots, R_n\}$ with a probability of $\{P_1^q, \ldots, P_n^q\}$ respectively. Next, for every image I_i in our test set, it predicts $R_1, \ldots, R_n$ with a confidence of $\{P_1^i, \ldots, P_n^i\}$. We calculate a matching score between an image with the query as $\sum_{j=1}^{n} P_j^q * P_j^i$. We also compare our model with Visual Phrases' detectors [6].

Results. SIFT [47] and GIST [46] descriptors perform poorly with a median rank of 54 and 68 (Table 5) because they simply measure structural similarity between images. CNN [44] descriptors capture object-level information and performs better with a median rank of 20. Our method captures the visual relationships present in the query image, which is important for high quality image retrieval, improving with a median rank of 4. When queried using an image of a "person riding a horse" (Figure 7), SIFT returns images that are visually similar but are not semantically relevant. CNN retrieves one image that contains a horse and one that contains both a man and a horse but neither of them capture the relationship: "person riding a horse". Visual Phrases and our model are able to detect the relationship ⟨person - ride - horse⟩ and perform better.

6 Conclusion

We proposed a model to detect multiple visual relationships in a single image. Our model learned to detect thousands of relationships even when there were

Algorithm 1 Training Algorithm

input: training set of images with annotated ⟨`subject - predicate - object`⟩ relationships annotated
Train object detectors on images using RCNN [43]
Train predicate classifier on images using VGG [44]
Initialize $f(\mathbf{W})$ (Eq. 2) with word vectors for objects using *word2vec*() [7]
repeat
    Compute the visual appearance model $V(\Theta)$ (Eq. 1)
    Compute relationship semantic distance to build $K(\mathbf{W})$ (Eq. 4)
    Compute the likelihood score $L(\mathbf{W})$ (Eq. 5)
    Backpropagate and optimize $\{\Theta, \mathbf{W}\}$ (Eq. 7) using stochastic gradient descent
until $\{\Theta, \mathbf{W}\}$ have converged
output: $\{\Theta, \mathbf{W}\}$

very few training examples. We learned the visual appearance of objects and predicates and combined them to predict relationships. To finetune our predictions, we utilized a language prior that mapped similar relationships together – outperforming previous state of the art [6] on the visual phrases dataset [6] as well as our dataset. We also demonstrated that our model can be used for zero shot learning of visual relationships. We introduced a new dataset with $37,993$ relationships that can be used for further benchmarking. Finally, by understanding visual relationships, our model improved content based image retrieval.

7 Supplementary Material

7.1 Training Algorithm

While we describe the theory and training procedure in the main text of this paper (Section 4.1), we include an algorithm box (Algorithm 1) to explain our training procedure in an alternate format.

7.2 Mean Average Precision on Visual Relationship Detection

As discussed in our paper, mean average precision (mAP) is a pessimistic evaluation metric for visual relationship detection because our dataset does not exhaustively annotate every possible relationship between two pairs of objects. For example, consider the case when our model predicts that a ⟨`person - next to - bicycle`⟩ when the ground truth annotation is ⟨`person - push - bicycle`⟩. In such a case, the prediction is not incorrect but would be penalized by mAP. However, to facilitate future comparisons against our model using this dataset, we report the mAP scores in Table 6.

We see a similar trend in the mAP scores as we did with the recall @ 50 and recall @ 100 values. The Visual Phrases [6] and Joint CNN baselines along with (Ours - L only) perform poorly on all three tasks: phrase, relationship and predicate detection. The visual only model (Ours - V only) improved upon

Table 6: mAP results for visual relationship detection (Section 5.1).

	Phrase Detection	Relationship Detection	Predicate Detection
Visual Phrases [6]	0.03	-	0.71
Joint CNN [44]	0.05	0.04	1.02
Ours - V only	0.93	0.84	6.42
Ours - L only	0.08	0.08	8.94
Ours - V + naive FC	1.21	1.19	11.05
Ours - V + L only	1.74	1.32	16.31
Ours - V + L + Reg.	1.78	1.40	17.95
Ours - V + L + K	**2.07**	**1.52**	**29.47**

Table 7: mAP results for zero-shot visual relationship detection (Section 5.2).

	Phrase Detection	Relationship Detection	Predicate Detection
Ours - V only	0.92	1.03	2.13
Ours - L only	0.00	0.00	3.31
Ours - V + L only	1.97	2.30	4.45
Ours - V + L + K	**2.89**	**3.01**	**5.52**

these results by leveraging the visual appearances of objects to aid it's predicate detection. Our complete model (Ours - V + L + K) achieves a mAP of 1.52 on relationship predication since it is penalized for missing annotations. However, it still performs better than all the other ablated models. It also attains a 29.47 mAP on predicate detection, demonstrating that our model learns to recognize predicates from one another.

7.3 Mean Average Precision on Zero-shot Learning

Similar to the previous section, we also include the mAP scores for zero shot learning in Table 7. Again, we see that the the inclusion of K() allows our model to levearage similar relationships to improve zero shot learning in all three experiments.

7.4 Human Evaluation on our Dataset

We ran an experiment to evaluate the human performance on our dataset. We randomly selecting 1000 pairs of objects from the dataset and then asked humans on Amazon Mechanical Turk to decide which of the 70 predicates were correct for each pair. We found that humans managed a 98.1% recall @ 50 and 96.4% mAP. This demonstrates that while this task is easy for humans, Visual Relationship Detection is still a hard unsolved task.

Acknowledgements . Our work is partially funded by an ONR MURI grant.

References

1. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2) (2010) 303–338
2. ZHOU12, G., Zhang, M., Ji, D.H., Zhu, Q.: Tree kernel-based relation extraction with context-sensitive structured parse tree information. EMNLP-CoNLL 2007 (2007) 728
3. GuoDong, Z., Jian, S., Jie, Z., Min, Z.: Exploring various knowledge in relation extraction. In: Proceedings of the 43rd annual meeting on association for computational linguistics, Association for Computational Linguistics (2005) 427–434
4. Culotta, A., Sorensen, J.: Dependency tree kernels for relation extraction. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics (2004) 423
5. Socher, R., Huval, B., Manning, C.D., Ng, A.Y.: Semantic compositionality through recursive matrix-vector spaces. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics (2012) 1201–1211
6. Sadeghi, M.A., Farhadi, A.: Recognition using visual phrases. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 1745–1752
7. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
8. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D.A., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
9. Mensink, T., Gavves, E., Snoek, C.G.: Costa: Co-occurrence statistics for zero-shot classification. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE (2014) 2441–2448
10. Salakhutdinov, R., Torralba, A., Tenenbaum, J.: Learning to share visual appearance for multiclass object detection. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 1481–1488
11. Ladicky, L., Russell, C., Kohli, P., Torr, P.H.: Graph cut based inference with co-occurrence statistics. In: Computer Vision–ECCV 2010. Springer (2010) 239–253
12. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in context. In: Computer vision, 2007. ICCV 2007. IEEE 11th international conference on, IEEE (2007) 1–8
13. Galleguillos, C., Rabinovich, A., Belongie, S.: Object categorization using co-occurrence, location and appearance. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE (2008) 1–8
14. Galleguillos, C., Belongie, S.: Context based object categorization: A critical survey. Computer Vision and Image Understanding **114**(6) (2010) 712–722
15. Choi, M.J., Lim, J.J., Torralba, A., Willsky, A.S.: Exploiting hierarchical context on a large database of object categories. In: Computer vision and pattern recognition (CVPR), 2010 IEEE conference on, IEEE (2010) 129–136
16. Izadinia, H., Sadeghi, F., Farhadi, A.: Incorporating scene context and object layout into appearance modeling. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE (2014) 232–239

17. Fidler, S., Leonardis, A.: Towards scalable representations of object categories: Learning a hierarchy of parts. In: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE (2007) 1–8
18. Sivic, J., Russell, B.C., Efros, A., Zisserman, A., Freeman, W.T., et al.: Discovering objects and their location in images. In: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on. Volume 1., IEEE (2005) 370–377
19. Gould, S., Rodgers, J., Cohen, D., Elidan, G., Koller, D.: Multi-class segmentation with relative location prior. International Journal of Computer Vision **80**(3) (2008) 300–316
20. Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M., Schiele, B.: Translating video content to natural language descriptions. In: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE (2013) 433–440
21. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 17–24
22. Maji, S., Bourdev, L., Malik, J.: Action recognition from a distributed representation of pose and appearance. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 3177–3184
23. Gupta, A., Kembhavi, A., Davis, L.S.: Observing human-object interactions: Using spatial and functional compatibility for recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on **31**(10) (2009) 1775–1789
24. Yao, B., Fei-Fei, L.: Grouplet: A structured image representation for recognizing human and object interactions. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 9–16
25. Ramanathan, V., Li, C., Deng, J., Han, W., Li, Z., Gu, K., Song, Y., Bengio, S., Rossenberg, C., Fei-Fei, L.: Learning semantic relationships for better action retrieval in images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 1100–1109
26. Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R., Darrell, T., Saenko, K.: Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE (2013) 2712–2719
27. Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., Pinkal, M.: Grounding action descriptions in videos. Transactions of the Association for Computational Linguistics **1** (2013) 25–36
28. Thomason, J., Venugopalan, S., Guadarrama, S., Saenko, K., Mooney, R.: Integrating language and vision to generate natural language descriptions of videos in the wild. In: Proceedings of the 25th International Conference on Computational Linguistics (COLING), August. (2014)
29. Yao, J., Fidler, S., Urtasun, R.: 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, IEEE (2012) 702–709
30. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Baby talk: Understanding and generating image descriptions. In: Proceedings of the 24th CVPR, Citeseer (2011)
31. Zitnick, C.L., Parikh, D., Vanderwende, L.: Learning the visual interpretation of sentences. In: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE (2013) 1681–1688

32. Gupta, A., Davis, L.S.: Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In: Computer Vision–ECCV 2008. Springer (2008) 16–29
33. Kumar, M.P., Koller, D.: Efficiently selecting regions for scene understanding. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 3217–3224
34. Russell, B.C., Freeman, W.T., Efros, A., Sivic, J., Zisserman, A., et al.: Using multiple segmentations to discover objects and their extent in image collections. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Volume 2., IEEE (2006) 1605–1614
35. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: Generating sentences from images. In: Computer Vision–ECCV 2010. Springer (2010) 15–29
36. Berg, A.C., Berg, T.L., Daume III, H., Dodge, J., Goyal, A., Han, X., Mensch, A., Mitchell, M., Sood, A., Stratos, K., et al.: Understanding and predicting importance in images. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 3562–3569
37. Hoiem, D., Efros, A.A., Hebert, M.: Putting objects in perspective. International Journal of Computer Vision **80**(1) (2008) 3–15
38. Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J., et al.: From captions to visual concepts and back. arXiv preprint arXiv:1411.4952 (2014)
39. Chang, A.X., Savva, M., Manning, C.D.: Semantic parsing for text to 3d scene generation. ACL 2014 (2014) 17
40. Schuster, S., Krishna, R., Chang, A., Fei-Fei, L., Manning, C.D.: Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In: Proceedings of the Fourth Workshop on Vision and Language (VL15). (2015)
41. Choi, W., Chao, Y.W., Pantofaru, C., Savarese, S.: Understanding indoor scenes using 3d geometric phrases. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE (2013) 33–40
42. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. In: International Journal of Computer Vision. (2016)
43. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Computer Vision and Pattern Recognition. (2014)
44. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
45. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. Pattern Analysis and Machine Intelligence, IEEE Transactions on **34**(11) (2012) 2189–2202
46. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision **42**(3) (2001) 145–175
47. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2) (2004) 91–110