Image Retrieval using Scene Graphs

Justin Johnson[1], Ranjay Krishna[1], Michael Stark[2], Li-Jia Li[3,4], David A. Shamma[3],
Michael S. Bernstein[1], Li Fei-Fei[1]

[1]Stanford University, [2]Max Planck Institute for Informatics, [3]Yahoo Labs, [4]Snapchat

Abstract

This paper develops a novel framework for semantic image retrieval based on the notion of a scene graph. Our scene graphs represent objects ("man", "boat"), attributes of objects ("boat is white") and relationships between objects ("man standing on boat"). We use these scene graphs as queries to retrieve semantically related images. To this end, we design a conditional random field model that reasons about possible groundings of scene graphs to test images. The likelihoods of these groundings are used as ranking scores for retrieval. We introduce a novel dataset of 5,000 human-generated scene graphs grounded to images and use this dataset to evaluate our method for image retrieval. In particular, we evaluate retrieval using full scene graphs and small scene subgraphs, and show that our method outperforms retrieval methods that use only objects or low-level image features. In addition, we show that our full model can be used to improve object localization compared to baseline methods.

1. Introduction

Retrieving images by describing their contents is one of the most exciting applications of computer vision. An ideal system would allow people to search for images by specifying not only objects ("man", "boat") but also structured *relationships* ("man on boat") and *attributes* ("boat is white") involving these objects. Unfortunately current systems fail for these types of queries because they do not utilize the structured nature of the query, as shown in Fig. 1.

To solve this problem, a computer vision system must explicitly represent and reason about the *objects*, *attributes*, and *relationships* in images, which we refer to as *detailed semantics*. Recently Zitnick et al. have made important steps toward this goal by studying *abstract scenes* composed of clip-art [71, 72, 22]. They show that perfect recognition of detailed semantics benefits image understanding and improves image retrieval.

Bringing this level of semantic reasoning to *real-world scenes* would be a major leap forward, but doing so involves two main challenges: *(1)* interactions between objects in a scene can be highly complex, going beyond simple pairwise relations, and *(2)* the assumption of a closed universe where all classes are known beforehand does not hold.

(a) Results for the query on a popular image search engine.

(b) Expected results for the query.

Figure 1: Image search using a complex query like "man holding fish and wearing hat on white boat" returns unsatisfactory results in (a). Ideal results (b) include correct *objects* ("man", "boat"), *attributes* ("boat is white") and *relationships* ("man on boat").

In order to address these challenges, this paper proposes a novel framework for detailed semantic image retrieval, based on a conditional random field (CRF [36]) model of visual scenes. Our model draws inspiration from recent work in computer graphics that uses graph-based formulations to compare [20] and generate [7] scenes. We use the notion of a *scene graph* to represent the detailed semantics of a scene.

Our scene graphs capture the detailed semantics of visual scenes by explicitly modeling objects, attributes of objects, and relationships between objects. Our model performs semantic image retrieval using scene graphs as queries. Replacing textual queries with scene graphs allows our queries to describe the semantics of the desired image in precise detail without relying on unstructured text. This formulation is related to a number of methods for object and scene recognition using context [25, 13]. But by using scene graphs, we can model multiple modes of interaction between pairs of objects while traditional CRF models are more restricted, and encode a fixed relation given two nodes (e.g. think of

"dog [eating, or playing, or being on the right of] a keyboard" in a scene graph versus. dog *on the right of* a keyboard in a CRF).

Specifically, our contributions are:

- We introduce a CRF model (Sect. 5) for semantic image retrieval using scene graph (Sect. 3) queries. We show that our model outperforms baseline models that reason only about objects, and simple content-based image retrieval methods based on low-level visual features (Sect. 6). This addresses challenge (1) above.
- We introduce a novel dataset[1] (Sect. 4) of $5,000$ human-annotated scene graphs grounded to images that use an open-world vocabulary to describe images in great detail, addressing challenge (2) above.

We set out to demonstrate the importance and utility of modeling detailed semantics using a scene graph representation for an image retrieval task. But as our experiments demonstrate, an advantage of this representation is its ability to offer deeper and detailed insights into images in a general framework. Our model can perform semantically meaningful image retrieval based on an entire scene (Sect. 6.1), or only parts of scenes (Sect. 6.2), and localizes specific objects (Sect. 6.3). This differentiates our intention and system from traditional content-based image retrieval work, which is not the focus of our paper.

2. Related Work

Image retrieval. Content-based image retrieval methods typically use low-level visual feature representations [50, 6], indexing [11, 69, 27, 28, 59], efficient sub-image search [38], object-category based retrieval [5, 62, 4], and other methods of modeling image similarity [52] to retrieve images based on their contents.

Text-based scene representations. There has been much recent interest in models that can jointly reason about images and natural language descriptions, ranging from Flickr tags [40] over sentences with canonical structure [35, 45, 72] to unaligned text corpora [57, 31, 58, 68]. While these models achieve impressive results in scene classification and object recognition using only weak supervision, they are typically limited in terms of expressiveness.

In contrast, our scene graphs are a structured representation of visual scenes. Each node is explicitly grounded in an image region, avoiding the inherent referential uncertainty of text-based representations. We currently use strong supervision in the form of a crowd-sourced data set (Sect. 4), but ultimately envision a system that learns novel scene graphs via active learning, like NEIL [8] or LEVAN [14].

Structured scene representations. More structured representations of visual scenes explicitly encode certain types of properties, such as attributes [19, 17, 16, 49, 37], object co-occurrence [44], or spatial relationships between pairs of objects [9, 13, 56, 10]. Our scene graphs generalize these representations, since they allow us to express each of them in a unified way (Sect. 3). Concretely, our CRF (Sect. 5) learns models for particular relations between pairs of objects, similar in spirit to [25] or [66]. However, we consider a much larger, open vocabulary of objects and relationships.

Graph-structured representations have attained widespread use in computer graphics to efficiently represent compositional scenes [1]. Fisher et al. [21] use graph kernels to compare 3D scenes, and Chang et al. [7] generate novel 3D scenes from natural language descriptions using a scene graph representation. Parse graphs obtained in scene parsing [70, 65] are typically the result of applying a grammar designed for a particular domain (such as indoor scenes [70]), in contrast to our generic scene graphs.

Recent work by Lin et al. [41] constructs semantic graphs from text queries using hand-defined rules to transform parse trees and uses these semantic graphs to retrieve videos in the context of autonomous driving. Their system is constrained to the six object classes from KITTI [23] while our system uses an open-world vocabulary; our scene graphs also tend to be more complex than their semantic graphs.

Zitnick et al. have studied detailed semantics using abstract scenes built from clip-art aligned to a corpus of sentences [71, 72, 22]. Our paper extends this work as follows: first, we explicitly address real-world scenes, replacing the idealistic setting of perfect object and attribute recognition [71] by uncertain detections. Second, our scene graphs go beyond pairwise relations: we model and discover meaningful higher-order interactions between multiple objects and attributes (these sub-graphs can be seen as a generalization of visual phrases [56, 10]). Third, our CRF for scene graph grounding (Sect. 5) can ground a query scene graph to an unannotated test image.

Real-world scene datasets. Datasets have been a driving force of computer vision algorithms, ranging from classification [64] to object recognition and segmentation [18, 12, 55, 15]. More recently, there has been a shift away from iconic images toward datasets depicting objects in the context of entire scenes, e.g., SUN09 [9], PASCAL-Context [46], and the COCO [42] datasets. Our novel dataset of real-world scene graphs advances in this direction by adding attributes and relationships. In contrast to previous work, the current version of our dataset with $5,000$ images focuses on depth rather than breadth (Sect. 4), resulting in a level of semantic detail that has not been achieved before. While this level of detail is similar in spirit to the Lotus Hill dataset [67], our dataset is based on an open universe assumption and is freely available.

[1] Available at the first author's website.

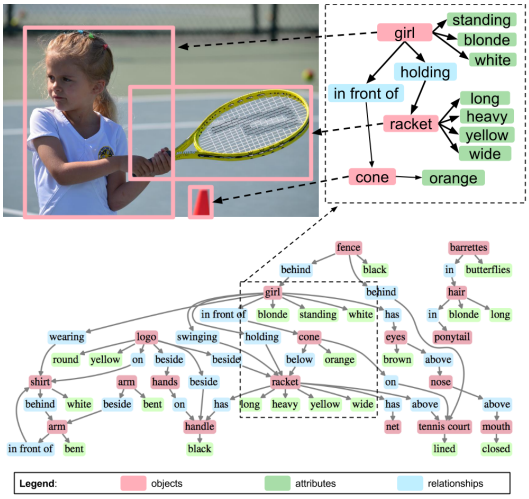

Figure 2: An example of a scene graph (bottom) and a grounding (top). The scene graph encodes objects ("girl"), attributes, ("girl is blonde"), and relationships ("girl holding racket"). The grounding associates each object of the scene graph to a region of an image. The image, scene graph, and grounding are drawn from our *real-world scene graphs* dataset (Sect. 4).

3. Scene Graphs

To retrieve images containing particular semantic contents, we need a formalized way of representing the contents of a scene. This representation must be powerful enough to describe the rich variety of scenes that can exist, without being too cumbersome. To this end, we define two abstractions: a *scene graph*, which is a way of describing a scene, and a *scene graph grounding*, which is a concrete association of a scene graph to an image.

3.1. Definition

A *scene graph* is a data structure that describes the contents of a scene. A scene graph encodes object instances, attributes of objects, and relationships between objects.

This simple formulation is powerful enough to describe visual scenes in great detail because it places no restriction on the types of objects, attributes, and relationships that can be represented. Fig. 2 (bottom) shows an example of a scene graph. In this example we see that object instances may be people ("girl"), places ("tennis court"), things ("shirt"), or parts of other objects ("arm"). Attributes can describe color ("cone is orange"), shape ("logo is round"), and pose ("arm is bent"). Relationships can encode geometry ("fence behind girl"), actions ("girl swinging racket"), and object parts ("racket has handle").

Formally, given a set of object classes $\mathcal{C}$, a set of attribute types $\mathcal{A}$, and a set of relationship types $\mathcal{R}$, we define a scene graph G to be a tuple $G = (O, E)$ where $O = \{o_1, \ldots, o_n\}$ is a set of objects and $E \subseteq O \times \mathcal{R} \times O$ is a set of edges. Each object has the form $o_i = (c_i, A_i)$ where $c_i \in \mathcal{C}$ is the class of the object and $A_i \subseteq \mathcal{A}$ are the attributes of the object.

3.2. Grounding a scene graph in an image

A scene graph on its own is not associated to an image; it merely describes a scene that could be depicted by an image. However a scene graph can be *grounded* to an image by associating each object instance of the scene graph to a region in an image. Fig. 2 (top) shows an example of part of a scene graph grounded to an image.

Formally, we represent an image by a set of candidate bounding boxes B. A grounding of a scene graph $G = (O, E)$ is then a map $\gamma : O \to B$. For ease of notation, for $o \in O$ we frequently write $\gamma(o)$ as γ_o.

Given a scene graph and an image, there are many possible ways of grounding the scene graph to the image. In Sect. 5 we formulate a method for determining the best grounding of a scene graph to an image.

3.3. Why scene graphs?

An obvious alternative choice for representing the content of scenes is natural language. However, in order to represent visual scenes at the level of detail shown in Fig. 2, a full paragraph of description would be necessary:

A blonde white girl is standing in front of an orange cone on a lined tennis court and is holding a long heavy yellow wide racket that has a black handle. The girl is wearing a white shirt; there is a bent arm in front of the shirt and another bent arm beside the first. There is a round yellow logo on the shirt, and the logo is beside hands that are on the handle of the racket. There is a black fence behind the girl, and the girl has brown eyes above a closed mouth. There are butterflies barrettes in long blonde hair, and the hair is in a ponytail.

To make use of such a description for image retrieval, we would need to resolve co-references in the text [53, 30, 39], perform relationship extraction to convert the unstructured text into structured tuples [47], and ground the entities of the tuples into regions of the image described by the text [33]. Such pipelines are challenging even in constrained settings [33], and would not scale to text of the detail shown above.

We can avoid these complexities by working directly with grounded scene graphs. We find that with careful user interface design, non-expert workers can quickly construct grounded scene graphs of arbitrary complexity. Details can be found in Sec. 4 and in our supplementary material.

4. Real-World Scene Graphs Dataset

To use scene graphs as queries for image retrieval, we need many examples of scene graphs grounded to images. To our knowledge no such dataset exists. To this end, we introduce a novel dataset of *real-world scene graphs*.

4.1. Data collection

We manually selected 5,000 images from the intersection of the YFCC100m [61] and Microsoft COCO [42] datasets, allowing our dataset to build upon rather than compete with these existing datasets.

	Full dataset	Experiments Sect. 6	COCO 2014 [42]	ILSVRC 2014 (Det) [54]	Pascal VOC [15]
Object classes	**6,745**	**266**	80	200	20
Attribute types	3,743	145	-	-	-
Relationship types	1,310	68	-	-	-
Object instances	93,832	69,009	**886,284**	534,309	27,450
Attribute instances	110,021	94,511	-	-	-
Relationship instances	112,707	109,535	-	-	-
Instances per obj. class	13.9	259.4	**11,087.5**	2,672.5	1,372
Instances per attr. type	29.4	651.8	-	-	-
Instances per rel. type	86.0	1,610.8	-	-	-
Objects per image	**18.8**	**13.8**	7.2	1.1	2.4
Attributes per image	22.0	18.9	-	-	-
Relationships per image	22.5	21.9	-	-	-
Attributes per object	1.2	1.0	-	-	-
Relationships per object	2.4	2.3	-	-	-

Table 1: Aggregate statistics for our *real-world scene graphs* dataset, for the full dataset and the restricted sets of object, attribute, and relationship types used in experiments.

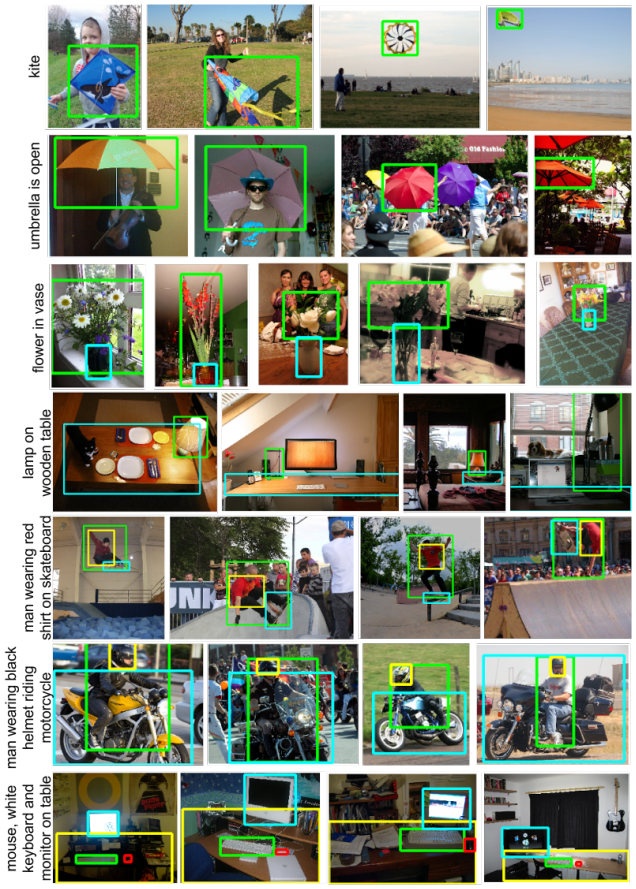

Figure 3: Examples of scene sub-graphs of increasing complexity (top to bottom) from our dataset, with attributes and up to 4 different objects.

For each of these images, we use Amazon's Mechanical Turk (AMT) to produce a human-generated scene graph.

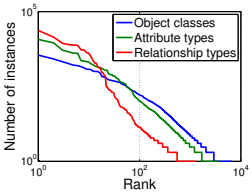

Figure 4: Objects, attributes and relations reveal a Zipf distribution when ordered by number of labeled instances.

For each image, three workers write (object, attribute) and (object, relationship, object) tuples using an open vocabulary to describe the image, and draw bounding boxes for all objects. Bounding boxes for objects are corrected and verified using a system similar to [60], and all tuples are verified by other workers. A detailed explanation of our data collection pipeline can be found in the supplementary material.

4.2. Analysis and statistics

Our full dataset of 5,000 images contains over 93,000 object instances, 110,000 instances of attributes, and 112,000 instances of relationships. For our experiments (Sect. 6), we consider only object classes and attribute types that appear at least 50 times in our training set and relationship types that occur at least 30 times in our training set. Even when we consider only the most common categories, as shown in Table 1, the dataset is still very rich with a mean of 13.8 objects, 18.9 attributes and 21.9 relationships per image.

A comparison of our dataset and other popular datasets can be found in Table 1. Compared to other datasets we prioritize detailed annotations of individual images over sheer quantity of annotated images. Our dataset contains significantly more labeled object instances per image than existing datasets, and also provides annotated attributes and relationships for individual object instances; these types of annotations are simply not available in other datasets. In addition, our decision to use an open vocabulary allows annotators to label the most meaningful features of each image instead of being constrained to a fixed set of predefined classes.

In contrast to previous work that looks at individual relations between pairs of objects in isolation [56], the deeply connected nature of our scene graph representation allows us to study object interactions of arbitrary complexity. Fig. 3 shows examples of scene-subgraphs that occur multiple times in our dataset, ranging from simple ("kite") to complex ("man on skateboard wearing red shirt"). These subgraphs also showcase the rich diversity expressed by our scene graphs. Attributes can for example encode object state ("umbrella is open"), material ("wooden table") and color ("red shirt", "black helmet"). Relationships can encode geometry ("lamp on table") as well as actions ("man riding motorcycle").

Fig. 5 uses our dataset to construct an *aggregate scene graph*, revealing the deeply connected nature of objects in the visual world. For example buses are frequently large and red, have black tires, are on streets, and have signs; signs

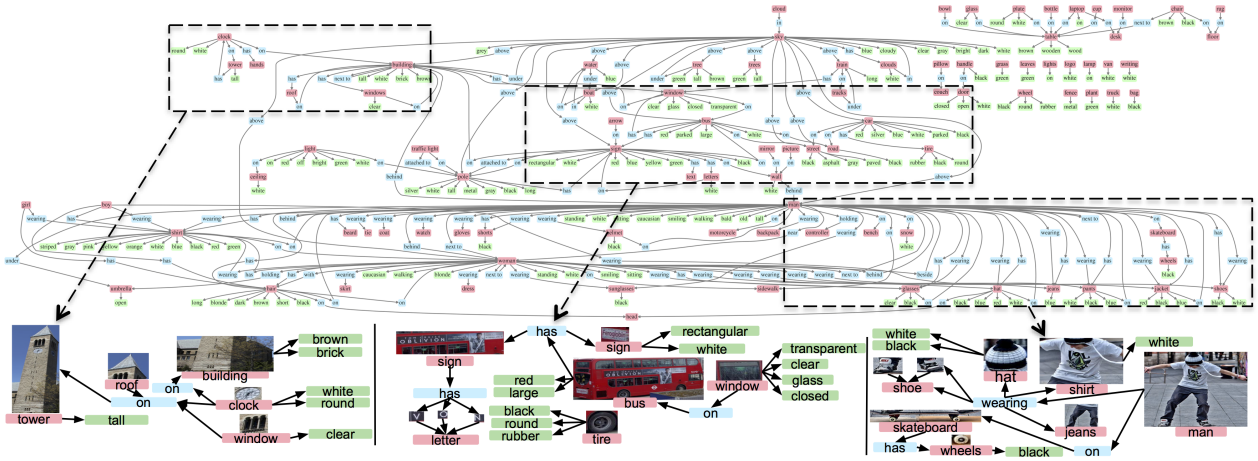

Figure 5: An *aggregate* scene graph computed using our entire dataset. We visualize the 150 most frequently occurring (object, relationship, object) and (object, attribute) tuples. We also provide 3 examples of scene graphs grounded in images that contribute to the sub-graphs within the dashed rectangles of the aggregated graph. Best viewed with magnification.

can also appear on walls, which often occur behind men, who often wear white shirts and jeans. The high density of the aggregate scene graph around the "man", "woman", "sky", and "building" nodes suggest that these objects are prominent elements of the visual world, but for different reasons: sky and building occur in nearly every image, while the attributes and relationships of people in scenes carries a huge semantic weight.

5. Image Retrieval by Scene Graph Grounding

We wish to use a scene graph as a query to retrieve images portraying scenes similar to the one described by the graph. To do so, we need to measure the agreement between a query scene graph and an unannotated test image. We assume that this agreement can be determined by examining the best possible grounding of the scene graph to the image.

To this end we construct a conditional random field (CRF [36]) that models the distribution over all possible groundings. We perform maximum a posteriori (MAP) inference to find the most likely grounding; the likelihood of this MAP solution is taken as the score measuring the agreement between the scene graph and the image.

5.1. CRF formulation

Reusing notation from Sect. 3, let $G = (O, E)$ be a scene graph, B be a set of bounding boxes in an image, and γ a grounding of the scene graph to the image. Each object $o \in O$ gives rise to a variable γ_o in the CRF, where the domain of γ_o is B and setting $\gamma_o = b \in B$ corresponds to grounding object o in the scene graph to box b in the image. We model the distribution over possible groundings as

$$P(\gamma \mid G, B) = \prod_{o \in O} P(\gamma_o \mid o) \prod_{(o,r,o') \in E} P(\gamma_o, \gamma_{o'} \mid o, r, o'). \quad (1)$$

We use Bayes' rule to rewrite the term $P(\gamma_o \mid o)$ as $P(o \mid \gamma_o)P(\gamma_o)/P(o)$. Assuming uniform priors over bounding boxes and object classes, $P(\gamma_o)$ and $P(o)$ are constants and can be ignored when performing MAP inference. Therefore our final objective has the form

$$\gamma^* = \arg\max_{\gamma} \prod_{o \in O} P(o \mid \gamma_o) \prod_{(o,r,o') \in E} P(\gamma_o, \gamma_{o'} \mid o, r, o'). \quad (2)$$

Unary potentials. The term $P(o \mid \gamma_o)$ in Equation 2 is a unary potential modeling how well the appearance of the box γ_o agrees with the known object class and attributes of the object o. If $o = (c, A)$ then we decompose this term as

$$P(o \mid \gamma_o) = P(c \mid \gamma_o) \prod_{a \in A} P(a \mid \gamma_o). \quad (3)$$

The terms $P(c \mid \gamma_o)$ and $P(a \mid \gamma_o)$ are simply the probabilities that the bounding box γ_o shows object class c and attribute a. To model these probabilities, we use R-CNN [24] to to train detectors for each of the $|\mathcal{C}| = 266$ and $|\mathcal{A}| = 145$ object classes and attribute types. We apply Platt scaling [51] to convert the SVM classification scores for each object class and attribute into probabilities.

Binary potentials. The term $P(\gamma_o, \gamma_{o'} \mid o, r, o')$ in Equation 2 is a binary potential that models how well the pair of bounding boxes $\gamma_o, \gamma_{o'}$ express the tuple (o, r, o'). Let $\gamma_o = (x, y, w, h)$ and $\gamma_{o'} = (x', y', w', h')$ be the coordinates of the bounding boxes in the image. We extract features $f(\gamma_o, \gamma_{o'})$ encoding their relative position and scale:

$$f(\gamma_o, \gamma_{o'}) = \Big((x - x')/w,\ (y - y')/h,\ w'/w,\ h'/h\Big) \quad (4)$$

Suppose that the objects o and o' have classes c and c' respectively. Using the training data, we train a Gaussian mixture model (GMM) to model $P(f(\gamma_o, \gamma_{o'}) \mid c, r, c')$. If there are fewer than 30 instances of the tuple (c, r, c') in the training data then we instead fall back on an object agnostic model $P(f(\gamma_o, \gamma_{o'}) \mid r)$. In either case, we use Platt scaling to convert the value of the GMM density function evaluated at $f(\gamma_o, \gamma_{o'})$ to a probability $P(\gamma_o, \gamma_{o'} \mid o, r, o')$.

5.2. Implementation details

We compared the performance of several methods for generating candidate boxes for images, including Objectness [2], Selective Search (SS [63]), and Geodesic Object Proposals (GOP [32]). We found that SS achieves the highest object recall on our dataset; however we use GOP for all experiments as it provides the best trade-off between object recall ($\approx 70\%$ vs $\approx 80\%$ for SS) and number of regions per image (632 vs 1720 for SS). We perform approximate inference using off-the-shelf belief propagation [3].

6. Experiments

We perform image retrieval experiments using two types of scene graphs as queries. First, we use full ground-truth scene graphs as queries; this shows that our model can effectively make sense of extremely precise descriptions to retrieve images. Second, we jump to the other end of the query complexity spectrum and use extremely simple scene graphs as queries; this shows that our model is flexible enough to retrieve relevant images when presented with more open-ended and human-interpretable queries.

In addition, we directly evaluate the groundings found by our model and show that our model is able to take advantage of scene context to improve object localization.

Setup. We randomly split our dataset into $4,000$ training images and $1,000$ test images. Our final vocabulary for object classes and attribute types is selected by picking all terms mentioned at least 50 times in the training set, and our vocabulary for relationship types is the set of relationships appearing at least 30 times in the training set. Statistics of our dataset when restricted to this vocabulary are shown in the second column of Table. 1. In our experiments we compare the following methods:

- **SG-obj-attr-rel**: Our model as described in Sect. 5. Includes unary object and attribute potentials and binary relationship potentials.
- **SG-obj-attr**: Our model, using only object and attribute potentials.
- **SG-obj**: Our model, using only object potentials. Equivalent to R-CNN [24] since the object class potentials are rescaled R-CNN detection scores.
- **CNN** [34]: L_2 distance between the last layer features extracted using the reference model from [29].
- **GIST** [48]: L_2 distance between the GIST descriptors of the query image and each test image (see Sect. 6.2).
- **SIFT** [43]: See Sect. 6.2.
- **Random**: A random permutation of the test images.

6.1. Full scene graph queries

We evaluate the performance of our model using the most complex scene graphs available to us – full human-generated scene graphs from our dataset. As argued in Sect. 3, these large scene graphs roughly correspond to a paragraph of text describing an image in great detail. Examples of query scene graphs and their rough textual equivalents are shown in the top half of Fig. 6.

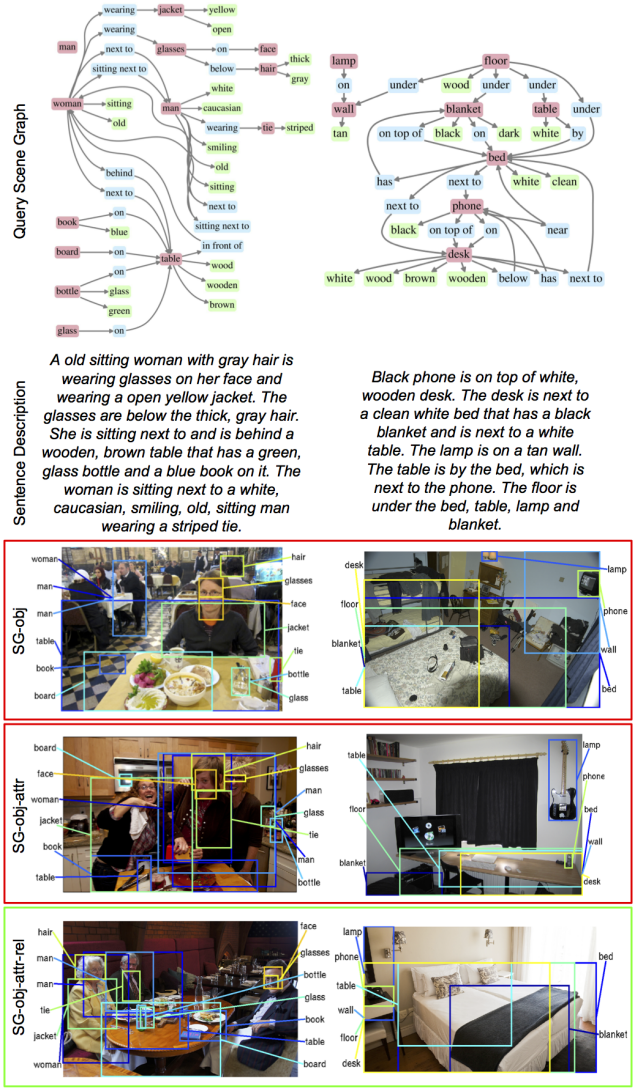

Figure 6: Example results for retrieval using full scene-graph queries (Sect. 6.1). Top: Example query graphs. Middle: Rough textual equivalents of the query scene graphs. Bottom: Top-1 retrieval results with groundings for our 3 methods. In both cases SG-obj-attr-rel succeeds in ranking the correct image at rank 1.

Concretely, we select an image I_q and its associated human-annotated scene graph G_q from our test set. We use the graph G_q as a query to rank all of the test images, and record the rank of the query image I_q. We repeat this process using 150 randomly selected images from our test set and evaluate the ranking of the query image over all 150 trials. Table 2 (a) gives the results in the form of recall at rank

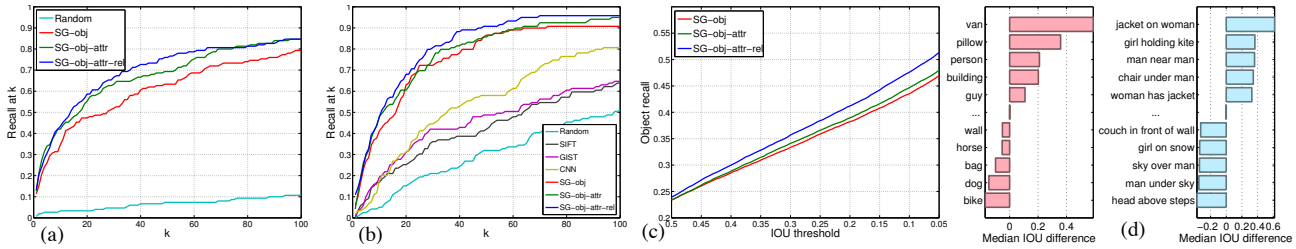

Figure 7: (a) Retrieval performance for entire scenes, (b) for partial scenes. (c) Object localization performance for entire scenes. (d) Increase in localization performance of our full model SG-obj-attr-rel vs SG-obj for individual objects (left) and objects participating in a relation (right). In (d), positive values indicate the SG-obj-attr-rel performs better than SG-obj.

		Rand	SIFT [43]	GIST [48]	CNN [34]	SG-obj [24]	SG-obj-attr	SG-obj-attr-rel
(a)	Med r	420	-	-	-	28	17.5	**14**
	R@1	0	-	-	-	0.113	0.127	**0.133**
	R@5	0.007	-	-	-	0.260	**0.340**	0.307
	R@10	0.027	-	-	-	0.347	0.420	**0.433**
(b)	Med r	94	64	57	36	17	12	**11**
	R@1	0	0	0.008	0.017	0.059	0.042	**0.109**
	R@5	0.034	0.084	0.101	0.050	0.269	0.294	**0.303**
	R@10	0.042	0.168	0.193	0.176	0.412	**0.479**	**0.479**
(c)	Med IoU	-	-	-	-	0.014	0.026	**0.067**
	R@0.1	-	-	-	-	0.435	0.447	**0.476**
	R@0.3	-	-	-	-	0.334	0.341	**0.357**
	R@0.5	-	-	-	-	0.234	0.234	**0.239**

Table 2: Quantitative results in entire scene retrieval ((a), Sect. 6.1), partial scene retrieval ((b), Sect. 6.2), and object localization ((c), Sect. 6.3).

k (higher is better) and median rank of the query image I_q (lower is better). Fig. 7 (a) plots the recall over k. Note that the GIST, SIFT, and CNN baselines are meaningless here, as they would always rank the query image highest.

Results. In Table 2 (a), we observe that the detailed semantics encoded in our scene graphs greatly increases the performance for entire scene retrieval. Compared to SG-obj, SG-obj-attr-rel decreases the median rank of the query image from 28 by half to 14. Recall at k shows similar results, where SG-obj-attr-rel increases recall over SG-obj by 2% (R@1), 4.7% (R@5), and 8.6% (R@10), respectively. This performance improvement increases for larger values of k (Fig.7 (a), blue vs red curve), to around 15% at 30. SG-obj-attr outperforms SG-obj, indicating that attributes are useful, and is in turn outperformed by SG-obj-attr-rel.

Fig. 6 shows corresponding qualitative results: on the left, our full model successfully identifies and localizes the "old woman with a jacket, sitting next to a sitting man with a striped tie". On the right, even though some objects are misplaced, SG-obj-attr-rel is able to correctly place the "dark blanket on the bed", while SG-obj-attr incorrectly grounds the blanket to a dark region of the test image.

6.2. Small scene graph queries

We have shown that our model is able to handle the complexities of full scene graph queries. Here we show that our model can also be used to retrieve meaningful results given simpler, more human-interpretable scene graph queries.

Specifically, we mine our dataset for re-occurring scene subgraphs containing two objects, one relationship, and one or two attributes, such as "sitting man on bench" and "smiling man wearing hat." We retain only subgraphs that occur at least 5 times in our test set, resulting in 598 scene subgraphs. We randomly selected 119 to be used as queries. For each subgraph query, we find the set of test images $I_1, \ldots, I_\ell$ that include it. We hold out I_1 from the test set and use the subgraph to rank the remaining 999 test images.

For the baseline methods we use the image I_1 rather than the graph to rank the test images. For the SIFT baseline, for each SIFT descriptor from I_1, we compute its 100 nearest neighbors among the descriptors from all other test images, and sort the test images by the number of these neighbors they contain. For the GIST and CNN baselines, we rank the test images based on the L_2 distance between the descriptor for I_1 and the descriptor for the test image.

We adapt the metrics from [26]; specifically, for each method we report the median rank of the highest ranked true positive image $I_2, \ldots, I_\ell$ across all queries and the recall at k for various values of k. Results are shown in Table 2 (b) and Fig. 7 (b). Examples of retrieved images can be seen in Fig. 8 (a,b), for the two queries mentioned above.

Results. In Table 2 (b), we see that our full model SG-obj-attr-rel again outperforms SG-obj by a large margin, reducing the median rank from 17 to 11. SG-obj-attr comes close, with a median rank of 12. In terms of recall at k, SG-obj-attr-rel again dominates, improving over SG-obj by 5% (R@1), 3.4% (R@5), and 6.7% (R@10), respectively. This gap increases up to around 12% at 40 (Fig. 7 (b)).

Qualitatively, Fig. 8 (a) shows that SG-obj-attr-rel retrieves correct results for "sitting man on bench." In comparison, the first result returned by SG-obj and SG-obj-attr contains both a man and a bench that are correctly localized, but the man is not sitting on the bench. Fig. 8 (b) shows that although SG-obj-attr-rel retrieves incorrect results for "smiling man wearing hat", it fails gracefully by returning images of a smiling woman wearing a hat, a man wearing a hat who is not smiling, and a smiling man wearing a helmet.

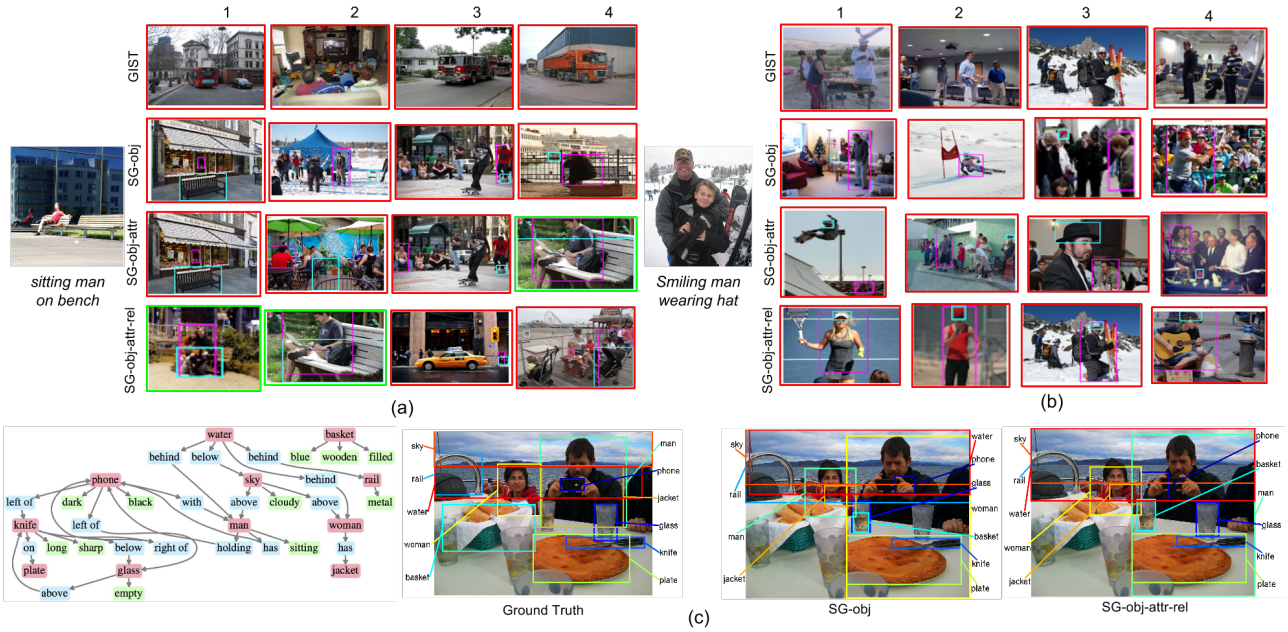

Figure 8: Top-*4* retrieval results returned by different methods using two different partial scene graph queries (a, b). Differences in fully automatic scene graph grounding when applying these methods to a particular test image (c).

6.3. Object localization

We have shown that our scene graph representation improves image retrieval results; here, we show that it can also aid in localizing individual objects. For each image I and its corresponding ground-truth (GT) scene graph G from our test set, we use each model to generate a grounding of G to I. For each object in G, we compute the intersection over union (IoU) between its GT position in the image and its position in the grounding generated by our model.

Fig. 8 (c) gives an example scene graph and its computed grounding under SG-obj-attr-rel and SG-obj. SG-obj labels "man" as "woman" and vice versa, but SG-obj-attr-rel is able to correct this error. Note that the scene graph does not specify any direct relationships between the man and the woman; instead, SG-obj-attr-rel must rely on the relationships between the two people and other objects in the scene ("man holding phone", "woman has jacket").

To quantitatively compare our models, we report the median IoU across all object instances in all test images, (Med IoU) and the fraction of object instances with IoU above various thresholds (IoU@t). Table 2 shows that SG-obj-attr-rel outperforms both SG-obj and SG-obj-attr on all metrics. Interestingly, comparing SG-obj-attr-rel to SG-obj shows a nearly five-fold increase in median IoU. The generally low values for median IoU highlight the difficulty of the automatic scene graph grounding task. Fig. 7 (c) shows that SG-obj-attr-rel performs particularly well compared to the baseline models at IoU thresholds below 0.5.

To gain more insight into the performance of our model, for each object instance in the test set we compute the difference in IoU with the GT between the grounding found by SG-obj-attr-rel and the grounding found by SG-obj. For each object class that occurs at least 4 times in the test set, we compute the median difference in IoU between the two methods, and perform the same analysis for (object, relationship, object) tuples, where for each such tuple in the test set we compute the mean IoU of the objects referenced by the tuple. The top and bottom 5 object classes and tuples are visualized in Fig. 7 (d).

This figure suggests that the context provided by SG-obj-attr-rel helps to localize rare objects (such as "van" and "guy") and objects with large appearance variations (such as "pillow" and "building"). SG-obj-attr-rel also gives large gains for tuples with well-defined spatial constraints ("jacket on woman", "chair under man"). Tuples encoding less well-defined spatial relationships ("man under sky", "girl on snow") may suffer in performance as the model penalizes valid configurations not seen at training time.

7. Conclusion

In this paper, we have used scene graphs as a novel representation for detailed semantics in visual scenes, and introduced a novel dataset of scene graphs grounded to real-world images. We have used this representation and dataset to construct a CRF model for semantic image retrieval using scene graphs as queries. We have shown that this model outperforms methods based on object detection and low-level visual features. We believe that semantic image retrieval is one of many exciting applications of our scene graph representation and dataset, and hope that more will follow.

8. Acknowledgments

This work has been partially supported by the Brown Institute for Media Innovation, the Max Planck Center for Visual Computing and Communication, and a National Science Foundation award IIS-1351131. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPUs used for this research. We would also like to thank Yuke Zhu, Jon Krause, Serena Yeung, and the anonymous reviewers for their constructive feedback.

References

[1] Openscenegraph. 2

[2] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2189–2202, 2012. 6

[3] B. Andres, T. Beier, and J. H. Kappes. OpenGM: A C++ library for discrete graphical models. *arXiv preprint arXiv:1206.0111*, 2012. 6

[4] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *Computer Vision–ECCV 2010*, pages 663–676. Springer, 2010. 2

[5] A. Bergamo, L. Torresani, and A. Fitzgibbon. Picodes: Learning a compact code for novel-category recognition. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2088–2096. 2011. 2

[6] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang. Spatial-bag-of-features. In *CVPR '10. IEEE Conference on Computer Vision and Pattern Recognition, 2010.*, 2010. 2

[7] A. X. Chang, M. Savva, and C. D. Manning. Learning spatial knowledge for text to 3D scene generation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, 2014. 1, 2

[8] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1409–1416. IEEE, 2013. 2

[9] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *IEEE Conference on Computer VIsion and Pattern Recognition (CVPR)*, 2010. 2

[10] W. Choi, Y.-W. Chao, C. Pantofaru, and S. Savarese. Understanding indoor scenes using 3d geometric phrases. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013. 2

[11] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007. 2

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 2

[13] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *International Journal of Computer Vision*, 95(1):1–12, 2011. 1, 2

[14] S. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, 2014. 2

[15] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 4

[16] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2352–2359. IEEE, 2010. 2

[17] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, 2009. 2

[18] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *CVPR-WS*, 2004. 2

[19] V. Ferrari and A. Zisserman. Learning visual attributes. In *NIPS*, 2007. 2

[20] M. Fisher, M. Savva, and P. Hanrahan. Characterizing structural relationships in scenes using graph kernels. In *ACM SIGGRAPH 2011 papers*, SIGGRAPH '11, pages 34:1–34:12. ACM, 2011. 1

[21] M. Fisher, M. Savva, and P. Hanrahan. Characterizing structural relationships in scenes using graph kernels. In *ACM SIGGRAPH 2011 papers*, SIGGRAPH '11, 2011. 2

[22] D. F. Fouhey and C. L. Zitnick. Predicting object dynamics in scenes. CVPR, 2014. 1, 2

[23] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2

[24] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition*, 2014. 5, 6, 7

[25] A. Gupta and L. S. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *Computer Vision–ECCV 2008*, pages 16–29. Springer, 2008. 1, 2

[26] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Res.(JAIR)*, 47:853–899, 2013. 7

[27] H. Jegou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2011. 2

[28] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2012. 2

[29] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolu-

tional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 6

[30] D. Jurafsky and J. H. Martin. *Speech & language processing*. Pearson Education India, 2000. 3

[31] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *CoRR*, abs/1406.5679, 2014. 2

[32] P. Krähenbühl and V. Koltun. Geodesic object proposals. In *Computer Vision–ECCV 2014*, pages 725–739. Springer, 2014. 6

[33] J. Krishnamurthy and T. Kollar. Jointly learning to parse and perceive: Connecting natural language to the physical world. *TACL*, 1:193–206, 2013. 3

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 6, 7

[35] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1601–1608. IEEE, 2011. 2

[36] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, 2001. 1, 5

[37] C. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot learning of object categories. *PAMI*, 2013. 2

[38] C. H. Lampert. Detecting objects in large image collections and videos by efficient subimage retrieval. In *Computer Vision, 2009 IEEE 12th International Conference on*, 2009. 2

[39] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34. Association for Computational Linguistics, 2011. 3

[40] L.-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, 2009. 2

[41] D. Lin, S. Fidler, C. Kong, and R. Urtasun. Visual semantic search: Retrieving videos via complex textual queries. In *CVPR*, 2014. 2

[42] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014. 2, 3, 4

[43] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 6, 7

[44] T. Mensink, E. Gavves, and C. G. Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *CVPR*, 2014. 2

[45] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daumé III. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics, 2012. 2

[46] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[47] F. Niu, C. Zhang, C. Ré, and J. W. Shavlik. Deepdive: Web-scale knowledge-base construction using statistical learning and inference. In *VLDS*, pages 25–28, 2012. 3

[48] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001. 6, 7

[49] D. Parikh and K. Grauman. Relative attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 503–510. IEEE, 2011. 2

[50] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *CVPR*, 2010. 2

[51] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*, 1999. 5

[52] D. Qin, C. Wengert, and L. Van Gool. Query adaptive similarity for large scale object retrieval. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013. 2

[53] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. Manning. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics, 2010. 3

[54] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge, 2014. 4

[55] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008. 2

[56] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1745–1752. IEEE, 2011. 2, 4

[57] R. Socher and L. Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 966–973. IEEE, 2010. 2

[58] R. Socher, Q. Le, C. Manning, and A. Ng. Grounded compositional semantics for finding and describing images with sentences. In *NIPS Deep Learning Workshop*, 2013. 2

[59] H. Stewenius and S. H. G. J. Pilet. Size matters: Exhaustive geometric verification for image retrieval. In *ECCV*. 2

[60] H. Su, J. Deng, and L. Fei-Fei. Crowdsourcing annotations for visual object detection. In *AAAI Technical Report, 4th Human Computation Workshop*, 2012. 4

[61] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015. 3

[62] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *European Conference on Computer Vision (ECCV)*, pages 776–789, Sept. 2010. 2

[63] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013. 6

[64] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010. 2

[65] J. Yang, B. Price, S. Cohen, and M.-H. Yang. Context driven scene parsing with attention to rare classes. In *CVPR*, 2014. 2

[66] B. Yao, A. Khosla, and L. Fei-Fei. Classifying actions and measuring action similarity by modeling the mutual context of objects and human poses. *a) A*, 1(D2):D3, 2011. 2

[67] Z. Yao, X. Yang, and S. Zhu. Introduction to a large scale general purpose groundtruth dataset: methodology, annotation tool, and benchmarks. In *EMMCVPR*, 2007. 2

[68] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 2

[69] X. Zhang, Z. Li, L. Zhang, W.-Y. Ma, and H.-Y. Shum. Efficient indexing for large scale visual search. In *Computer Vision, 2009 IEEE 12th International Conference on*, 2009. 2

[70] Y. Zhao and S.-C. Zhu. Scene parsing by integrating function, geometry and appearance models. In *CVPR*, 2013. 2

[71] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3009–3016. IEEE, 2013. 1, 2

[72] C. L. Zitnick, D. Parikh, and L. Vanderwende. Learning the visual interpretation of sentences. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1681–1688. IEEE, 2013. 1, 2