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.

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...
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.

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.
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 \( C \), 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 \( c_i \in C \) 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 of 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 \( \gamma : O \rightarrow B \). For ease of notation, for \( o \in O \) we frequently write \( \gamma(o) \) as \( \gamma_o \).

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 [33, 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 [38], 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.
### 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.

<table>
<thead>
<tr>
<th></th>
<th>Full dataset</th>
<th>Experiments Sect. 6</th>
<th>COCO 2014</th>
<th>ILSVRC 2014 (Det)</th>
<th>Pascal VOC 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Object classes</strong></td>
<td>6,745</td>
<td>266</td>
<td>80</td>
<td>200</td>
<td>20</td>
</tr>
<tr>
<td><strong>Attribute types</strong></td>
<td>3,743</td>
<td>145</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Relationship types</strong></td>
<td>1,310</td>
<td>68</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Object instances</strong></td>
<td>93,832</td>
<td>69,009</td>
<td>886,284</td>
<td>534,309</td>
<td>27,450</td>
</tr>
<tr>
<td><strong>Attribute instances</strong></td>
<td>110,021</td>
<td>94,511</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Relationship instances</strong></td>
<td>112,707</td>
<td>109,535</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Instances per obj. class</strong></td>
<td>13.9</td>
<td>29.4</td>
<td>11,087.5</td>
<td>2,672.5</td>
<td>1,372</td>
</tr>
<tr>
<td><strong>Instances per attr. type</strong></td>
<td>29.4</td>
<td>651.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Instances per rel. type</strong></td>
<td>86.0</td>
<td>1,610.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Objects per image</strong></td>
<td>18.8</td>
<td>13.8</td>
<td>7.2</td>
<td>1.1</td>
<td>2.4</td>
</tr>
<tr>
<td><strong>Attributes per image</strong></td>
<td>22.0</td>
<td>18.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Relationships per image</strong></td>
<td>22.5</td>
<td>21.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Attributes per object</strong></td>
<td>1.2</td>
<td>1.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Relationships per object</strong></td>
<td>2.4</td>
<td>2.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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. 4 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...
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, 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 \( \gamma \) a grounding of the scene graph to the image. Each object \( o \in O \) gives rise to a variable \( \gamma_o \) in the CRF, where the domain of \( \gamma_o \) 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

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

\[
\gamma^* = \arg \max_{\gamma} \prod_{o \in O} P(o \mid \gamma_o) \prod_{(o, r, o') \in E} P(\gamma_o, \gamma_{o'} \mid o, r, o').
\] (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 \( \gamma_o \) agrees with the known object class and attributes of the object \( o \). If \( o = (c, A) \) then we decompose this term as

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

The terms \( P(c \mid \gamma_o) \) and \( P(a \mid \gamma_o) \) are simply the probabilities that the bounding box \( \gamma_o \) shows object class \( c \) and attribute \( a \). To model these probabilities, we use R-CNN [24] to to train detectors for each of the \(|A| = 145 \) 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:

\[
f(\gamma_o, \gamma_{o'}) = \left( (x - x')/w, (y - y')/h, w'/w, h'/h \right)
\] (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 (≈ 70% vs ≈ 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.

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
Recall at $k$

Random $\text{SG−obj−obj−attr}$ $\text{SG−obj−obj−attr−rel}$

Increase in localization performance of our full model $\text{SG−obj−attr−rel}$ vs $\text{SG−obj}$ for individual objects (left) and objects participating in a relation (right). In (d), positive values indicate the $\text{SG−obj−attr−rel}$ performs better than $\text{SG−obj}$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Rand I</th>
<th>SIFT [33]</th>
<th>GIST [89]</th>
<th>CNN [34]</th>
<th>SG−obj</th>
<th>SG−obj−attr</th>
<th>SG−obj−attr−rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Med r</td>
<td>420</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>28</td>
<td>17.5</td>
<td>14</td>
</tr>
<tr>
<td>R@1</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.113</td>
<td>0.121</td>
<td>0.133</td>
</tr>
<tr>
<td>R@5</td>
<td>0.007</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.260</td>
<td>0.340</td>
<td>0.307</td>
</tr>
<tr>
<td>R@10</td>
<td>0.027</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.347</td>
<td>0.420</td>
<td>0.433</td>
</tr>
<tr>
<td>(b) Med r</td>
<td>94</td>
<td>64</td>
<td>57</td>
<td>36</td>
<td>17</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>R@1</td>
<td>0</td>
<td>0.034</td>
<td>0.084</td>
<td>0.101</td>
<td>0.050</td>
<td>0.269</td>
<td>0.294</td>
</tr>
<tr>
<td>R@5</td>
<td>0.042</td>
<td>0.168</td>
<td>0.193</td>
<td>0.176</td>
<td>0.412</td>
<td>0.479</td>
<td>0.479</td>
</tr>
<tr>
<td>R@10</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(c) Med IoU</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.014</td>
<td>0.026</td>
<td>0.067</td>
</tr>
<tr>
<td>R@1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.415</td>
<td>0.447</td>
<td>0.476</td>
</tr>
<tr>
<td>R@5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.334</td>
<td>0.341</td>
<td>0.357</td>
</tr>
<tr>
<td>R@10</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.234</td>
<td>0.234</td>
<td>0.239</td>
</tr>
</tbody>
</table>

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 $\text{SG−obj}$, $\text{SG−obj−attr−rel}$ decreases the median rank of the query image from 28 by half to 14. Recall at $k$ shows similar results, where $\text{SG−obj−attr−rel}$ increases recall over $\text{SG−obj}$ 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. $\text{SG−obj−attr−rel}$ outperforms $\text{SG−obj}$, indicating that attributes are useful, and is in turn outperformed by $\text{SG−obj−attr−rel}$.

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, $\text{SG−obj−attr−rel}$ is able to correctly place the “dark blanket on the bed”, while $\text{SG−obj−attr}$ 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 $I_1, \ldots, I_t$ 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 $I_2, \ldots, I_t$ 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 $\text{SG−obj−attr−rel}$ again outperforms $\text{SG−obj}$ by a large margin, reducing the median rank from 17 to 11. $\text{SG−obj−attr}$ comes close, with a median rank of 12. In terms of recall at $k$, $\text{SG−obj−attr−rel}$ again dominates, improving over $\text{SG−obj}$ 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 $\text{SG−obj−attr−rel}$ retrieves correct results for “sitting man on bench.” In comparison, the first result returned by $\text{SG−obj}$ and $\text{SG−obj−attr}$ 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 $\text{SG−obj−attr−rel}$ 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.
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


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