

## What are you talking about? Text-to-Image Coreference

Chen Kong<sup>1</sup> Dahua Lin<sup>3</sup> Mohit Bansal<sup>3</sup> Raquel Urtasun<sup>2,3</sup> Sanja Fidler<sup>2,3</sup>

<sup>1</sup>Tsinghua University, <sup>2</sup>University of Toronto, <sup>3</sup>TTI Chicago

kc10@mails.tsinghua.edu.cn, {dhlin,mbansal}@ttic.edu,{fidler,urtasun}@cs.toronto.edu

### Abstract

In this paper we exploit natural sentential descriptions of RGB-D scenes in order to improve 3D semantic parsing. Importantly, in doing so, we reason about which particular object each noun/pronoun is referring to in the image. This allows us to utilize visual information in order to disambiguate the so-called coreference resolution problem that arises in text. Towards this goal, we propose a structure prediction model that exploits potentials computed from text and RGB-D imagery to reason about the class of the 3D objects, the scene type, as well as to align the nouns/pronouns with the referred visual objects. We demonstrate the effectiveness of our approach on the challenging NYU-RGBD v2 dataset, which we enrich with natural lingual descriptions. We show that our approach significantly improves 3D detection and scene classification accuracy, and is able to reliably estimate the text-to-image alignment. Furthermore, by using textual and visual information, we are also able to successfully deal with coreference in text, improving upon the state-of-the-art Stanford coreference system [15].

### 1. Introduction

Imagine a scenario where you wake up late on a Saturday morning and all you want is for your personal robot to bring you a shot of bloody mary. You could say “It is in the upper cabinet in the kitchen just above the stove. I think it is hidden behind the box of cookies, which, please, bring to me as well.” For a human, finding the mentioned items based on this information should be an easy task. The description tells us that there are at least two cabinets in the kitchen, one in the upper part. There is also a stove and above it is a cabinet holding a box and the desired item should be behind it. For autonomous systems, sentential descriptions can serve as rich source of information. Text can help us parse the visual scene in a more informed way, and can facilitate for example new ways of active labeling and learning.

Understanding descriptions and linking them to visual content is fundamental to enable applications such as semantic visual search and human-robot interaction. Using language to provide annotations and guide an automatic

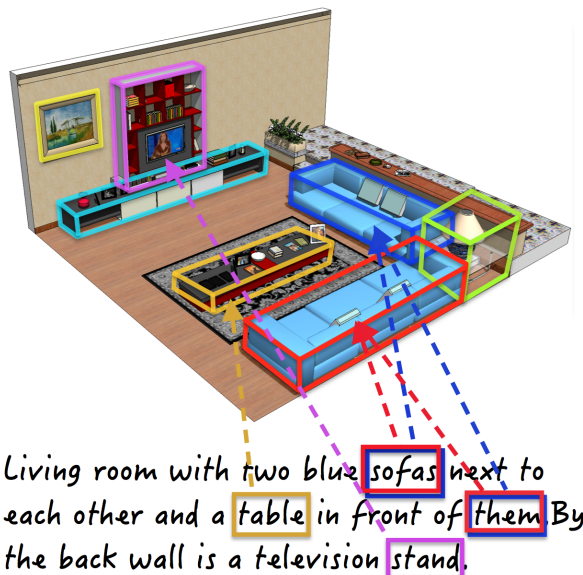


Figure 1. Our model uses lingual descriptions (a string of dependent sentences) to improve visual scene parsing as well as to determine which visual objects the text is referring to. We also deal with coreference within text (e.g., pronouns like “it” or “them”).

system is key for the deployment of such systems. To date, however, attempts to utilize more complex natural descriptions are rare. This is due to the inherent difficulties of both natural language processing and visual recognition, as well as the lack of datasets that contain such image descriptions linked to visual annotations (e.g., segmentation, detection).

Most recent approaches that employ text and images focus on generation tasks, where given an image one is interested in generating a lingual description of the scene [8, 12, 21, 2], or given a sentence, retrieving related images [29]. An exception is [9], which employed nouns and prepositions extracted from short sentences to boost the performance of object detection and semantic segmentation.

In this paper we are interested in exploiting natural lingual descriptions of RGB-D scenes in order to improve 3D object detection as well as to determine which particular object each noun/pronoun is referring to in the image. In order to do so, we need to solve the text to image alignment

problem, as illustrated with the arrows in Fig. 1. Towards this goal, we propose a holistic model that reasons jointly about the visual scene and as well as text that accompanies the image. Our model is a Markov Random Field (MRF) which reasons about the type of scene, 3D detection as well as to which visual concept each noun/pronoun refers to. We demonstrate the effectiveness of our approach in the challenging NYU-RGBD v2 dataset [27] which we enrich with natural lingual descriptions. Our model is able to significantly improve 3D detection and scene classification accuracy over the visual-only baseline. Furthermore, it successfully deals with the text to image alignment problem, as well as with coreference resolution in text, improving over the state-of-the-art Stanford coreference system [15].

## 2. Related Work

There has been substantial work in automatic captioning or description generation of images [12, 13, 14] and video [2]. Text has also been used as a form of weak supervision to learn visual models. In [35], representations for word meanings are learned from short video clips paired with sentences by learning the correspondence between words and video concepts in an unsupervised fashion. Ramanathan et al. [23] use textual descriptions of videos to learn action and role models. Matuszek et al. [19] jointly learn language and perception models for grounded attribute induction. In [10], prepositional relations and adjectives are used to learn object detection models from weakly labeled data. Natural descriptions have also recently been used as semantic queries for video retrieval [17].

Very little work has been devoted to exploiting text to improve semantic visual understanding beyond simple image classification [22], or tag generation [7, 3]. Notable exceptions are [16, 28], which exploit weak labels in the form of tags to learn models that reason about classification, annotation as well as segmentation. The work closest to us is [9], where short sentences are parsed into nouns and prepositions, which are used to generate potentials in a holistic scene model [34]. However, the approach does not identify which instance the sentence is talking about, and for example whenever a sentence mentions a “car”, the model boosts all car hypotheses in the image. In contrast, here we are interested in *aligning* nouns and pronouns in text with the referred objects in the image. Further, in our work we use complex natural descriptions composed of several sentences, as opposed to single sentences used in [9], where additional challenges such as coreference emerge.

Only a few datasets contain images and text. The UIUC dataset [8] augments a subset of PASCAL’08 with 3 independent sentences (each by a different annotator) on average per image, thus typically not providing very rich descriptions. Other datasets either do not contain visual labels (e.g., Im2Text [21]) or tackle a different problem, e.g., ac-

tion recognition. We expect the dataset we collect here to be of great use to both the NLP and vision communities.

## 3. Text to Image Alignment Model

Our input is an RGB-D image of an indoor scene as well as its multi-sentence description. Our goal is to jointly parse the 3D visual scene and the text describing the image, as well as to match the text to the visual concepts, performing text to image alignment. We frame the problem as the one of inference in a Markov Random Field (MRF) which reasons about the type of scene, 3D detection as well as for each noun/pronoun of interest which visual concept it correspond to. To cope with the exponentially many detection candidates we use bottom-up grouping to generate a smaller set of “objectness” cuboid hypothesis, and restrict the MRF to reason about those. In the following, we describe how we parse the text in Sec. 3.1, generate 3D object candidates in Sec 3.2, and explain our MRF model in Sec. 3.3.

### 3.1. Parsing Textual Descriptions

We extract part of speech tags (POS) of all sentences in a description using the Stanford POS Tagger for English language [31]. Type dependencies were obtained using [6].

**Nouns:** We extract nouns from the POS, and match them to the object classes of interest. In order to maximize the number of matched instances, we match nouns not only to the name of the class, but also to its synonyms and plural forms. We obtain the synonyms via WordNet as well as from our text ground-truth (GT). In particular, we add to our list of synonyms all words that were annotated as object class of interest in the training split of the text GT.

**Attributes:** To get attributes for a noun we look up all *amod* and *nsubj* that modify a noun of interest in Stanford’s parser dependency information. Here *amod* means an adjectival modifier, e.g., *amod(brown, table)*, and *nsubj* is a nominal subject, e.g. for “The table is brown” we have *nsubj(brown,table)*. The parser can handle multiple attributes per noun, which we exploit.

**Pronouns:** Since the sentences in our descriptions are not independent but typically refer to the same entity multiple times, we are faced with the so-called *coreference resolution* problem. For example, in “A table is in the room. Next to it is a chair.”, both *table* and *it* refer to the same object and thus form a coreference. To address this, we use the Stanford coreference system [15] to predict clusters of coreferent mentions. This is a state-of-the-art, rule-based and sieve-based system which works well on out-of-domain data. We generate multiple coreference hypotheses<sup>1</sup> in order to allow the MRF model to correct for some of the coreference mistakes using both textual and visual information.

<sup>1</sup>We ran [15] with different values of `maxdist` parameter which is the maximum sentence distance allowed between two mentions for resolution.

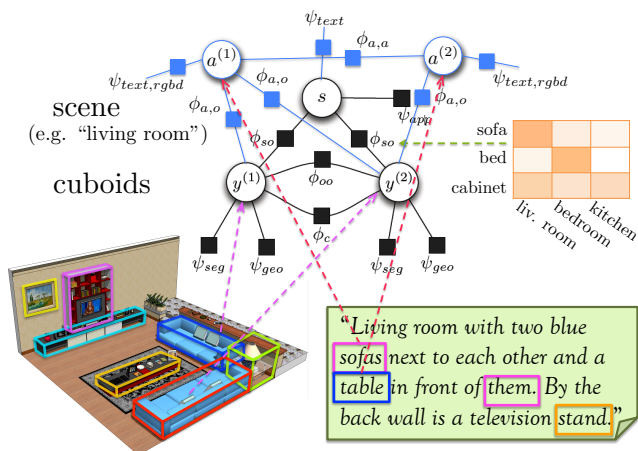


Figure 2. Our model. Black nodes and arrows represent visual information [18], blue are text and alignment related variables.

Pronouns are tagged as PRP by the Stanford parser. We take all PRP occurrences that are linked to a noun of interest in our coreference clusters. We extract attributes for each such pronoun just as what we do for a noun.

### 3.2. Visual Parsing

Our approach works with a set of object candidates represented as 3D cuboids and reasons about the class of each cuboid within our joint visual and textual MRF model. We follow [18] to get cuboid candidates by generating ranked 3D “objectness” regions that respect intensity and occlusion boundaries in 3D. To get the regions we use CPMC [5] extended to 3D [18]. Each region is then projected to 3D via depth and a cuboid is fit around it by requiring the horizontal faces to be parallel to the ground. In our experiments, we vary the number  $K$  of cuboid hypotheses per image.

### 3.3. Our Joint Visual and Textual Model

We define a Markov Random Field (MRF) which reasons about the type of scene, 3D detection as well as for each noun/pronoun of interest which visual concept it correspond to. More formally, let  $s \in \{1, \dots, S\}$  be a random variable encoding the type of scene, and let  $y_i \in \{0, \dots, C\}$ , with  $i = 1, \dots, K$ , be a random variable associated with a candidate cuboid, encoding its semantic class. Here  $y_i = 0$  indicates that the cuboid is a false positive. Let  $\mathcal{T}$  be the textual information and let  $\mathcal{I}$  be the RGB-D image evidence. For each noun/pronoun corresponding to a class of interest we generate an indexing random variable  $a_j \in \{0, \dots, K\}$  that *selects* the cuboid that the noun refers to. The role of variables  $\{a_j\}$  is thus to align text with visual objects. Note that  $a_j = 0$  means that there is no cuboid corresponding to the noun. This could arise in cases where our set of candidate cuboids does not contain the object that the noun describes, or the noun refers to an virtual object that is in fact not visible in the scene. For plural forms we generate as many  $a$  variables as the cardinality of the (pro)noun. Our

MRF energy sums the energy terms exploiting the image and textual information while reasoning about the text to image alignment. The graphical model is depicted in Fig. 2. We now describe the potentials in more detail.

#### 3.3.1 Visual Potentials

Our visual potentials are based on [18] and we only briefly describe them here for completeness.

**Scene Appearance:** To incorporate global information we define a unary potential over the scene label, computed by means of a logistic on top of a classifier score [33].

**Cuboid class potential:** We employ CPMC-o2p [4] to obtain detection scores. For the background class, we use a constant threshold. We use an additional classifier based on the segmentation potentials of [24], which classify superpixels. To compute our cuboid scores, we compute the weighted sum of scores for the superpixels falling within the convex hull of the cuboid in the image plane.

**Object geometry:** We capture object’s geometric properties such as height, width, volume, as well as its relation to the scene layout, e.g., distance to the wall. We train an SVM classifier with an RBF kernel for each class, and use the resultant scores as unary potentials for the candidate cubes.

**Semantic context:** We use two co-occurrence relationships: *scene-object* and *object-object*. The potential values are estimated by counting the co-occurrence frequencies.

**Geometric context:** We use two potentials to exploit the spatial relations between cuboids in 3D, encoding *close-to* and *on-top-of* relations. The potential values are simply the empirical counts for each relation.

#### 3.3.2 Text-to-Image Alignment Potentials

A sentence describing an object can carry rich information about its properties, as well as 3D relations within the scene. For example, a sentence “There is a wide wooden table by the left wall.” provides size and color information about the table as well as roughly where it can be found in the room. We now describe how we encode this in the model.

**Scene from Text:** In their description of the image, people frequently provide information about the type of scene. This can be either direct, e.g., “This image shows a child’s *bedroom*.”, or indirect “There are two single beds in this room”. We encode this with the following potential:

$$\phi_{scene}(s = u|\mathcal{T}) = \gamma_u, \quad (1)$$

where  $\gamma_u$  is the score of an SVM classifier (RBF kernel). To compute its features, we first gathered occurrence statistics for words in the training corpus and took only words that exceeded a minimum occurrence threshold. We did not use stop words such as *and* or *the*. We formed a feature vector with an 0/1 entry for each word, depending whether the

|             | mantel | counter | toilet | sink | bath tub | bed  | headb. | table | shelf | cabinet | sofa | chair | chest | refrig. | oven | microw. | blinds | curtain | board | monitor | printer |
|-------------|--------|---------|--------|------|----------|------|--------|-------|-------|---------|------|-------|-------|---------|------|---------|--------|---------|-------|---------|---------|
| % mentioned | 41.4   | 46.4    | 90.6   | 78.2 | 55.6     | 79.4 | 9.2    | 53.4  | 32.9  | 27.7    | 51.2 | 31.6  | 45.7  | 55.7    | 31.4 | 56.0    | 26.2   | 23.0    | 28.6  | 66.7    | 44.3    |
| # nouns     | 19     | 948     | 120    | 219  | 84       | 547  | 7      | 1037  | 240   | 1105    | 375  | 631   | 131   | 68      | 57   | 46      | 41     | 74      | 64    | 226     | 58      |
| # of coref  | 3      | 71      | 28     | 49   | 11       | 145  | 0      | 212   | 27    | 129     | 80   | 57    | 8     | 7       | 11   | 1       | 2      | 8       | 29    | 8       | 8       |

Table 1. **Dataset statistics.** **First row:** Number of times an object class is mentioned in text relative to all visual occurrences of that class. **Second row:** The number of noun mentions for each class. **Third row:** Number of coreference occurrences for each class.

description contained the word or not. For nouns referring to scene types we merged all synonym words (e.g., “dining room”, “dining area”) into one entry of the feature vector.

**Alignment:** This unary potential encodes the likelihood that the  $j$ -th relevant noun/pronoun is referring to the  $i$ -th cuboid in our candidate set. We use the score of a classifier,

$$\phi_{ao}^{\text{align}}(a_j = i | \mathcal{I}, \mathcal{T}) = \text{score}(a_j, i), \quad (2)$$

which exploits several visual features representing the  $i$ -th cuboid, such as its color, aspect ratio, distance to wall, and textual features representing information relevant to  $a_j$ ’s noun, such as “wide”, “brown”, “is close to wall”. On the visual side, we use 10 geometric features to describe each cuboid (as in Sec. 3.3.1), as well as the score of a color classifier for six main colors that people most commonly used in their descriptions (*white, blue, red, black, brown, bright*). To exploit rough class information we also use the cuboid’s segmentation score (Sec. 3.3.1). Since people typically describe more salient objects (bigger and image centered), we use the dimensions of the object’s 2D bounding box (a box around the cuboid projected to the image), and its  $x$  and  $y$  image coordinates as additional features. We also use the center of the cuboid in 3D to capture the fact that people tend to describe objects that are closer to the viewer.

On the text side, we use a class feature which is 1 for  $a_j$ ’s noun’s class and 0 for other object classes of interest. We also use a color feature, which is 1 if the noun’s adjective is a particular color, and 0 otherwise. To use information about the object’s position in the room, we extract 9 geometric classes from text: *next-to-wall, in-background, in-foreground, middle-room, on-floor, on-wall, left-side-room, right-side-room, in-corner-room*, with several synonyms for each class. We form a feature that is 1 if the sentence with  $a_j$  mentions a particular geometric class, and 0 otherwise. We train an SVM classifier (RBF kernel) and transform the scores via a logistic function (scale 1) to form the potentials.

**Size:** This potential encodes how likely the  $j$ -th relevant noun/pronoun refers to the  $i$ -th cuboid, given that it was mentioned to have a particular physical size in text:

$$\phi_{ao}^{\text{size}}(a_j = i | \mathcal{I}, \mathcal{T}, \text{size}(j) = sz) = \text{score}_{sz}(\text{class}(j), i) \quad (3)$$

We train a classifier for two sizes, *big* and *small*, using geometric features such as object’s width and height in 3D. We use several synonyms for both, including *e.g. double, king, single*, typical adjectives for describing beds. Note that size depends on object class: a “big cabinet” is not of the same physical size as a “big mug”. Since we do not have enough

training examples for (size,class) pairs, we add another feature to the classifier which is 1 if the noun comes from a group of big objects (e.g., bed, cabinet) and 0 otherwise.

**Text-cuboid Compatibility:** This potential ensures that the class of the cuboid that a particular  $a_j$  selects matches the  $a_j$ ’s noun. We thus form a noun-class compatibility potential between  $a_j$  and each cuboid:

$$\phi_{ao}^{\text{compat.}}(a_j = i, y_i = c) = \text{stats}(\text{class}(j), c) \quad (4)$$

When  $a_j$  corresponds to a pronoun we find a noun from its coreference cluster and take its class. We use empirical counts as our potential rather than a hard constraint as people confuse certain types of classes. This is aggravated by the fact that our annotators were not necessarily native speakers. For example people would say “chest” instead of “cabinet”. These errors are challenges of natural text.

**Variety:** Notice that our potentials so far are exactly the same for all  $a_j$  corresponding to the *same plural mention*, e.g., “Two *chairs* in the room.” However, since they are referring to different objects in the scene we encourage them to point to different cuboids via a pairwise potential:

$$\phi_{a,a}^{\text{variety}}(a_j, a_k | \mathcal{T}) = \begin{cases} -1 & \text{if } a_j = a_k \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

### 3.4. Learning and Inference

Inference in our model is NP-hard. We use distributed convex belief propagation (DCBP) [25] to perform approximate inference. For learning the weights for each potential in our MRF, we use the primal-dual method of [11], specifically, we use the implementation of [26].

We define the loss function as a sum of unary and pairwise terms. In particular, we define a 0-1 loss over the scene type, and a 0-1 loss over the cuboid detections [18]. For the assignment variables  $a_j$  we use a thresholded loss:

$$\Delta_a(a_j, \hat{a}_j) = \begin{cases} 1 & \text{if } \text{IOU}(a_j, \hat{a}_j) \leq 0.5 \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

which penalizes alignments with cuboids that have less than 50% overlap with the cuboids in the ground-truth alignments. For plural forms, the loss is 0 if  $a_j$  selects **any** of the cuboids in the noun’s ground-truth alignment. In order to impose diversity, we also use a pairwise loss requiring that each pair  $(a_j, a_k)$ , where  $a_j$  and  $a_k$  correspond to different object instances for a plural noun mention, needs to select two different cuboids:

$$\Delta_{\text{plural}}(a_j, a_k) = \begin{cases} 1 & \text{if } a_j = a_k \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

| # sent              | # words    | min # sent        | max sent      | min words | max words |
|---------------------|------------|-------------------|---------------|-----------|-----------|
| 3.2                 | 39.1       | 1                 | 10            | 6         | 144       |
| # nouns of interest | # pronouns | # scene mentioned | scene correct |           |           |
| 3.4                 | 0.53       | 0.48              | 83%           |           |           |

Table 2. Statistics per description.

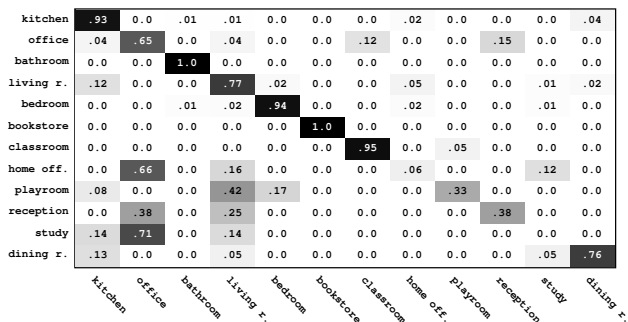


Figure 3. Scene classif. accuracy with respect to NYU annotation. We evaluate acc. only when a scene is *mentioned* in a description.

#### 4. RGB-D Dataset with Complex Descriptions

Having rich data is important in order to enable automatic systems to properly ground language to visual concepts. Towards this goal, we took the NYUv2 dataset [27] that contains 1449 RGB-D images of indoor scenes, and collected sentential descriptions for each image. We asked the annotators (MTurkers) to describe an image to someone who does not see it to give her/him a vivid impression of what the scene looks like. The annotators were only shown the image and had no idea of what the classes of interest were. For quality control, we checked all descriptions, and fixed those that were grammatically incorrect, while preserving the content. The collected descriptions go beyond current datasets where typically only a short sentence is available. They vary from one to ten sentences per annotator per image, and typically contain rich information and multiple mentions to objects. Fig. 6 shows examples.

We collected two types of ground-truth annotations. The first one is visual, where we linked the nouns and pronouns to the visual objects they describe. This gives us ground-truth alignments between text and images. We used in-house annotators to ensure quality. We took a conservative approach and labeled only the non-ambiguous referrals. For plural forms we linked the (pro)noun to multiple objects.

The second annotation is text based. Here, the annotators were shown **only text** and not the image, and thus had to make a decision based on the syntactic and semantic textual information alone. For all nouns that refer to the classes of interest we annotated which object class it is, taking into account synonyms. All other nouns were marked as *background*. For each noun we also annotated attributes (i.e., color and size) that refer to it. We also annotated co-referrals in cases where different words talk about the same entity by linking the head (representative) noun in a description to all its noun/pronoun occurrences. We annotated at-

|              | precision | recall | F-measure |
|--------------|-----------|--------|-----------|
| object class | 94.7%     | 94.2%  | 94.4%     |
| scene        | 85.7%     | 85.7%  | 85.7%     |
| color        | 64.2%     | 93.0%  | 75.9%     |
| size         | 55.8%     | 96.0%  | 70.6%     |

Table 3. Parser accuracy (based on Stanford’s parser [31])

| Method        | MUC       |        |       | B <sup>3</sup> |        |       |
|---------------|-----------|--------|-------|----------------|--------|-------|
|               | precision | recall | F1    | precision      | recall | F1    |
| Stanford [15] | 61.56     | 62.59  | 62.07 | 75.05          | 76.15  | 75.59 |
| Ours          | 83.69     | 51.08  | 63.44 | 88.42          | 70.02  | 78.15 |

Table 4. Co-reference accuracy of [15] and our model.

tributes for the linked pronouns as well. Our annotation was semi-automatic, where we generated candidates using the Stanford parser [31, 15] and manually corrected the mistakes. We used WordNet to generate synonyms.

We analyze our dataset next. Table 2 shows simple statistics: there are on average 3 sentences per description where each description has on average 39 words. Descriptions contain up to 10 sentences and 144 words. A pronoun belonging to a class of interest appears in every second description. Scene type is explicitly mentioned in half of the descriptions. Table 1 shows per class statistics, e.g. percentage of times a noun refers to a visual object with respect to the number of all visual objects of that class. Interestingly, a “toilet” is talked about 91% of times it is visible in a scene, while “curtains” are talked about only 23% of times. Fig. 4 shows size histograms for the mentioned objects, where size is the square root of the number of pixels which the linked object region contains. We separate the statistics into whether the noun was mentioned in the first, second, third, or fourth and higher sentence. An interesting observation is that the sizes of mentioned objects become smaller with the sentence ID. This is reasonable as the most salient (typically bigger) objects are described first. We also show a plot for sizes of objects that are mentioned more than once per description. We can see that the histogram is pushed to the right, meaning that people corefer to bigger objects more often. As shown in Fig. 5, people first describe the closer and centered objects, and start describing other parts of the scene in later sentences. Finally, in Fig. 3 we evaluate human scene classification accuracy against NYU ground-truth. We evaluate accuracy only when a scene is explicitly *mentioned* in a description. While “bathroom” is always a “bathroom”, there is confusion for some other scenes, e.g. a “playroom” is typically mentioned to be a “living room”.

#### 5. Experimental Evaluation

We test our model on the NYUv2 dataset augmented with our descriptions. For 3D object detection we use the same class set of 21 objects as in [18], where ground-truth has been obtained by robust fitting of cuboids around object regions projected to 3D via depth. For each image NYU also has a scene label, with 13 scene classes altogether.

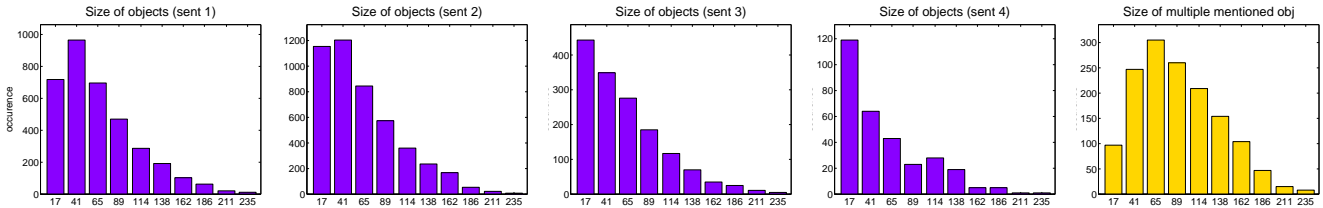


Figure 4. Statistics of sizes of described object regions give the sequential sentence number in a description.

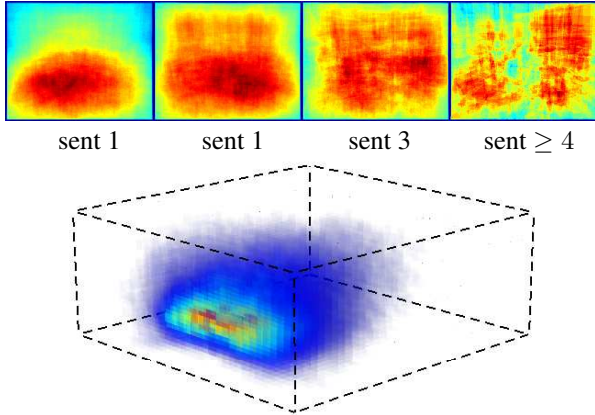


Figure 5. **Top:** Location statistics of mentioned objects given which sentence mentions them. **Bottom:** Statistics of mentioned objects in 3D. We translate each scene to have the ground at  $y = 0$ .

### 5.1. Performance metrics

**Cuboid:** An object hypotheses is said to be correct if it overlaps with a ground-truth object (of the same class) more than 50% IOU. Detector’s performance is then measured in terms of the F-measure averaged across the classes.

**Text-cuboid alignment:** We say that an alignment  $a_j = i$  is correct if IOU between the  $i$ -th candidate cuboid and the cuboid in  $a_j$ ’s GT alignment is higher than 50%. For plural nouns with several  $\{a_{j_k}\}$ , we take a union of all cuboids  $i_k$  in the inferred alignments:  $a_{j_k} = i_k$ . We compute IOU between this union and the union of cuboids in the GT alignment for this noun. We evaluate via F-measure. We use two metrics, one for all alignments in our GT, including nouns/pronouns, and one for only the alignment of nouns.

**Scene:** Accuracy is computed as the average along the diagonal of the confusion matrix.

**Text parser:** In order to understand the performance of our model, it is important to sanity check the performance of the Stanford-based **text** parser in our setting. The results are shown in Table 3 in terms of F-measure. We can see that while the object and scene performance is relatively good, parsing of the attributes is lagging behind.

**Coreference metrics:** We evaluate the predicted coreference clusters with MUC [32] and  $B^3$  [1] metrics. MUC is a link-based metric that measures how many predicted mention clusters need to be merged to cover the gold (GT) clusters.  $B^3$  is a mention-based metric which computes precision and recall for each mention as the overlap between its predicted and gold cluster divided by the size of the predicted and gold cluster, respectively. As shown by previ-

ous work [20], the MUC metric is somewhat biased towards large clusters, whereas the  $B^3$  metric is biased towards singleton clusters. We thus report both metrics simultaneously, so as to ensure a fair and balanced evaluation. To compute the metrics we need a matching between the GT and predicted mentions. We consider a GT and predicted mention to be matched if the GT’s head-word is contained in the prediction’s mention span.<sup>2</sup> All unmatched mentions (GT and predicted) are penalized appropriately by our metrics.<sup>3</sup>

### 5.2. Results

To examine the performance of our model in isolation, we first test it on the GT cuboids. In this way, the errors introduced in the detection stage are not considered. We measure the performance in terms of *classification accuracy*, i.e., the percentage of correctly classified objects. We consider various configurations of our model that incorporate different subsets of potentials, so as to analyze their individual effect. The results are presented in Table 5. By “+” we denote that we added a potential to the setting in the previous line. Here “ $a$  coref” denotes that we used the  $a$  variables also for the coreferred words (see Sec. 3.1) as opposed to just for the explicit noun mentions. We can observe that overall our improvement over the visual-only model is significant, 14.4% for scene classification and 6.4% for object detection. Further, the full model improves 2.3% over the unary-based alignment when evaluating nouns only (denoted with *align N* in the Table) as well as for both nouns and pronouns (denoted with *align N+P*).

For real cuboids, we test our performance by varying the number  $K$  of cuboid hypotheses per scene. The results are shown in Table 5 showing that performance in all tasks improves more than 5% over the baseline. The joint model also boosts the text to cuboid alignment performance over unary only (which can be considered as an intelligent text-visual baseline), most notably for  $K = 15$ . Performance for each class is shown in Table 6, showing a big boost in accuracy for certain classes. Note that for GT cuboids, Table 5 shows classification accuracy, while Table 6 shows F-measure. For real cuboids both tables show F-measure.

Note that our model can also re-reason about coreference in text by exploiting the visual information. This is done via

<sup>2</sup>This is because we start with “visual” head-words as the gold-standard mentions. In case of ties, we prefer mentions which are longer (if they have the same head-word) or closer to the head-word of the matching mention.

<sup>3</sup>We use a version of the  $B^3$  metric called  $B^3_{all}$ , defined by [30] that can handle (and suitably penalize) unmatched GT and predicted mentions.

|                   | Ground-truth cuboids |             |             |             | our cuboids ( $K = 8$ ) |             |             |             | our cuboids ( $K = 15$ ) |             |             |             |
|-------------------|----------------------|-------------|-------------|-------------|-------------------------|-------------|-------------|-------------|--------------------------|-------------|-------------|-------------|
|                   | scene                | object      | align. N    | align N+P   | scene                   | object      | align N     | align N+P   | scene                    | object      | align N     | align N+P   |
| [18]              | 58.1                 | 60.5        | –           | –           | 38.1                    | 44.3        | –           | –           | 54.4                     | 37.3        | –           | –           |
| random assign.    | –                    | –           | 15.8        | 15.3        | –                       | –           | 5.9         | 5.7         | –                        | –           | 4.7         | 4.6         |
| assign. to bckg.  | –                    | –           | 7.5         | 7.3         | –                       | –           | 7.7         | 7.5         | –                        | –           | 7.7         | 7.5         |
| + scene text      | 67.1                 | 60.7        | –           | –           | 62.4                    | 44.3        | –           | –           | 61.6                     | 37.3        | –           | –           |
| + $a$ unary       | 67.1                 | 60.6        | 51.4        | 49.7        | 62.5                    | 44.3        | 20.4        | 19.8        | 60.9                     | 37.3        | 20.4        | 19.7        |
| + $a$ size        | 67.1                 | 60.6        | 51.6        | 49.9        | 62.5                    | 44.3        | 20.4        | 19.7        | 60.9                     | 37.3        | 20.4        | 19.8        |
| + $a$ -to-cuboid  | 72.2                 | 66.8        | 53.2        | 51.5        | 67.0                    | <b>49.1</b> | <b>24.0</b> | 23.2        | <b>61.9</b>              | 43.5        | 24.8        | 24.0        |
| + $a$ coref       | 72.3                 | <b>67.1</b> | 53.6        | 51.8        | <b>67.1</b>             | 48.8        | 23.5        | 22.7        | <b>61.9</b>              | 43.4        | 24.9        | 24.1        |
| + $a$ variability | <b>72.5</b>          | 67.0        | <b>53.7</b> | <b>52.0</b> | <b>67.1</b>             | 48.7        | 23.5        | <b>22.8</b> | 61.6                     | <b>44.1</b> | <b>25.1</b> | <b>24.3</b> |

Table 5. Results with the baseline and various instantiations of our model. Here *assign N* means noun-cuboid assignment accuracy (F measure), and *assign N+P* where we also evaluate assignment of pronouns to cuboids.  $K$  means the number of cuboid candidates used.

the inferred  $\{a_j\}$  variables: each  $a_j$  in our MRF selects a cuboid which the corresponding (pro)noun talks about. We then say that different  $a$  variables selecting *the same cuboid* form a coreference cluster. Table 4 shows the accuracy of the Stanford’s coreference system [15] and our approach. Note that in our evaluation we discard (in Stanford’s and our output) all predicted clusters that contain *only non-gold* mentions. This is because our ground-truth contains coreference annotation only for 21 classes of interest, and we do not want to penalize precision outside of these classes, especially for the Stanford system which is class agnostic. Our results show that we can improve over a text-only state-of-the-art system [15] via a joint text-visual model. This is, to the best of our knowledge, the first result of this kind.

## 6. Conclusions

We proposed a holistic MRF model which reasons about 3D object detection, scene classification and alignment between text and visual objects. We showed a significant improvement over the baselines in challenging scenarios in terms of both visual scene parsing, text to image alignment and coreference resolution. In future work, we plan to employ temporal information as well, while at the same time reasoning over a larger set of objects, stuff and verb classes.

**Ack.:** Our work was partially supported by ONR N00014-13-1-0721.

## References

- [1] A. Bagga and B. Baldwin. Algorithms for scoring coreference chains. In *MUC-7 and LREC Workshop*, 1998. 6
- [2] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, L. Schmidt, J. Shangquan, J. Siskind, J. Waggoner, S. Wang, J. Wei, Y. Yin, and Z. Zhang. Video-in-sentences out. In *UAI*, 2012. 1, 2
- [3] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. In *JMLR*, 2003. 2
- [4] J. Carreira, R. Caseiroa, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV12*, 2012. 3
- [5] J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *TPAMI*, 2012. 3
- [6] M. de Marneffe, B. MacCartney, and C. Manning. Generating typed dependency parses from phrase structure parses. In *LREC*, 2006. 2
- [7] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002. 2
- [8] A. Farhadi, M. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences for images. In *ECCV*, 2010. 1, 2
- [9] S. Fidler, A. Sharma, and R. Urtasun. A sentence is worth a thousand pixels. In *CVPR*, 2013. 1, 2
- [10] A. Gupta and L. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*, 2008. 2
- [11] T. Hazan and R. Urtasun. A primal-dual message-passing algorithm for approximated large scale structured prediction. In *NIPS*, 2010. 4
- [12] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. Berg, and T. Berg. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, 2011. 1, 2
- [13] P. Kuznetsova, V. Ordonez, A. Berg, T. Berg, and Y. Choi. Collective generation of natural image descriptions. In *Association for Computational Linguistics (ACL)*, 2012. 2
- [14] P. Kuznetsova, V. Ordonez, A. Berg, T. Berg, and Y. Choi. Generalizing image captions for image-text parallel corpus. In *Association for Computational Linguistics (ACL)*, 2013. 2
- [15] H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916, 2013. 1, 2, 5, 7
- [16] L. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, 2009. 2
- [17] D. Lin, S. Fidler, C. Kong, and R. Urtasun. Visual semantic search: Retrieving videos via complex textual queries. In *CVPR*, 2014. 2
- [18] D. Lin, S. Fidler, and R. Urtasun. Holistic scene understanding for 3d object detection with rgb cameras. In *ICCV*, 2013. 3, 4, 5, 7, 8
- [19] C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox. A joint model of language and perception for grounded attribute learning. In *ICML*, 2012. 2
- [20] V. Ng. Supervised noun phrase coreference research: The first fifteen years. In *ACL*, 2010. 6

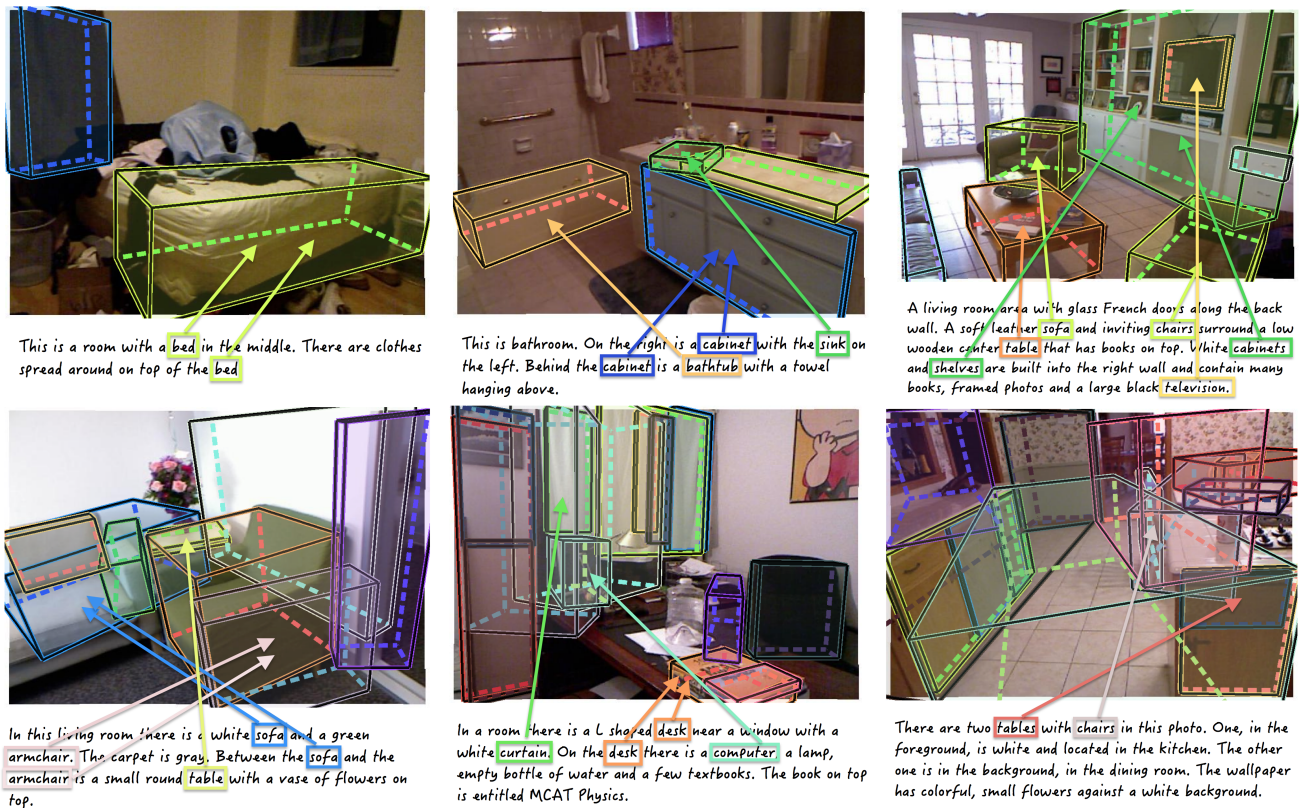


Figure 6. Qualitative results on text to visual object alignment.

|          |      | mantel      | count.      | toilet      | sink        | tub         | bed         | head.       | table       | shelf       | cabin.      | sofa        | chair       | chest       | refrig.     | oven        | micro.      | blinds      | curtain     | board       | monit.      | print.      | avg.        |
|----------|------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| GT       | [18] | 0.0         | 83.9        | 33.3        | 63.6        | 12.5        | 68.6        | 25.0        | 50.7        | <b>59.5</b> | 75.4        | 54.9        | 80.2        | <b>37.8</b> | 7.1         | 6.5         | 15.4        | 63.0        | 54.7        | 41.5        | 31.9        | 8.0         | 41.6        |
| cuboids  | ours | <b>16.7</b> | <b>84.7</b> | <b>80.0</b> | <b>81.8</b> | <b>33.3</b> | <b>81.7</b> | <b>28.1</b> | <b>66.3</b> | 58.6        | <b>76.0</b> | <b>61.5</b> | <b>83.2</b> | 35.6        | <b>35.7</b> | <b>9.7</b>  | <b>23.1</b> | <b>64.9</b> | <b>59.7</b> | <b>43.4</b> | <b>58.4</b> | <b>40.0</b> | <b>53.4</b> |
| our      | [18] | 100.0       | 50.3        | 53.5        | <b>45.6</b> | 0.0         | 56.7        | 34.8        | 45.4        | 54.3        | 50.8        | 55.6        | <b>50.9</b> | <b>50.6</b> | 26.1        | <b>20.0</b> | 45.5        | 49.5        | 42.9        | <b>47.1</b> | 50.0        | 0.0         | 44.3        |
| $K = 8$  | ours | 100.0       | <b>52.5</b> | <b>55.1</b> | 43.9        | 0.0         | <b>60.3</b> | <b>36.4</b> | <b>51.2</b> | <b>55.4</b> | <b>51.2</b> | <b>55.8</b> | 50.3        | 47.1        | <b>38.5</b> | 0.0         | <b>57.1</b> | <b>52.4</b> | <b>43.6</b> | 44.4        | <b>61.8</b> | <b>66.7</b> | <b>48.7</b> |
| our      | [18] | 50.0        | 41.2        | 45.4        | <b>43.0</b> | 20.0        | 51.8        | 30.3        | 37.5        | <b>45.8</b> | 43.9        | 46.5        | 45.8        | <b>38.4</b> | 22.7        | 10.0        | 40.0        | 45.9        | 41.4        | 39.3        | 43.9        | 0.0         | 37.3        |
| $K = 15$ | ours | 50.0        | <b>45.7</b> | <b>48.4</b> | 42.9        | 20.0        | <b>53.6</b> | <b>32.3</b> | <b>43.0</b> | 44.8        | <b>45.5</b> | <b>48.8</b> | <b>46.6</b> | 36.6        | <b>45.8</b> | <b>19.0</b> | <b>51.9</b> | <b>46.6</b> | <b>41.9</b> | <b>51.9</b> | <b>54.9</b> | <b>55.6</b> | <b>44.1</b> |
| our      | [18] | 0.0         | 36.4        | 44.6        | 37.0        | <b>11.1</b> | 46.2        | 25.5        | 29.4        | <b>39.6</b> | 40.0        | 37.8        | 40.3        | <b>32.5</b> | 13.7        | 7.1         | 34.3        | <b>44.1</b> | <b>38.4</b> | 26.1        | 43.0        | 0.0         | 29.9        |
| $K = 30$ | ours | <b>20.0</b> | <b>39.3</b> | <b>47.5</b> | <b>39.4</b> | 10.8        | <b>48.8</b> | <b>26.7</b> | <b>33.8</b> | 39.1        | <b>40.4</b> | <b>40.0</b> | <b>41.0</b> | 31.7        | <b>39.0</b> | <b>13.3</b> | <b>41.0</b> | <b>45.6</b> | 38.1        | <b>34.3</b> | <b>52.8</b> | <b>34.5</b> | <b>36.1</b> |

Table 6. Class-wise performance comparison between the baseline (visual only model) and our model.

- [21] V. Ordonez, G. Kulkarni, and T. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 1, 2
- [22] A. Quattoni, M. Collins, and T. Darrell. Learning visual representations using images with captions. In *CVPR*, 2007. 2
- [23] V. Ramanathan, P. Liang, and L. Fei-Fei. Video event understanding using natural language descriptions. In *ICCV*, 2013. 2
- [24] X. Ren, L. Bo, and D. Fox. Rgb-(d) scene labeling: Features and algorithms. In *CVPR*, 2012. 3
- [25] A. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Distributed message passing for large scale graphical models. In *CVPR*, 2011. 4
- [26] A. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Efficient structured prediction with latent variables for general graphical models. In *ICML*, 2012. 4
- [27] N. Silberman, P. Kohli, D. Hoiem, and R. Fergus. Indoor segmentation and support inference from rgb-d images. In *ECCV*, 2012. 2, 5
- [28] R. Socher and L. Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *CVPR*, 2010. 2
- [29] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPS*, 2012. 1
- [30] V. Stoyanov, N. Gilbert, C. Cardie, and E. Riloff. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *ACL/IJCNLP*, 2009. 6
- [31] K. Toutanova, D. Klein, and C. Manning. Feature-rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL*, 2003. 2, 5
- [32] M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. A model-theoretic coreference scoring scheme. In *MUC-6*, 1995. 6
- [33] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 3
- [34] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012. 2
- [35] H. Yu and J. M. Siskind. Grounded language learning from video described with sentences. In *ACL*, 2013. 2