A Corpus of Natural Language for Visual Reasoning

Alane Suhr[†], **Mike Lewis**[‡], **James Yeh**[†], **and Yoav Artzi**[†]

[†] Dept. of Computer Science and Cornell Tech, Cornell University, New York, NY 10044
{suhr, yoav}@cs.cornell.edu, jamesclyeh@gmail.com

[‡] Facebook AI Research, Menlo Park, CA 94025
mikelewis@fb.com

Abstract

We present a new visual reasoning language dataset, containing 92,244 pairs of examples of natural statements grounded in synthetic images with 3,962 unique sentences. We describe a method of crowdsourcing linguistically-diverse data, and present an analysis of our data. The data demonstrates a broad set of linguistic phenomena, requiring visual and set-theoretic reasoning. We experiment with various models, and show the data presents a strong challenge for future research.

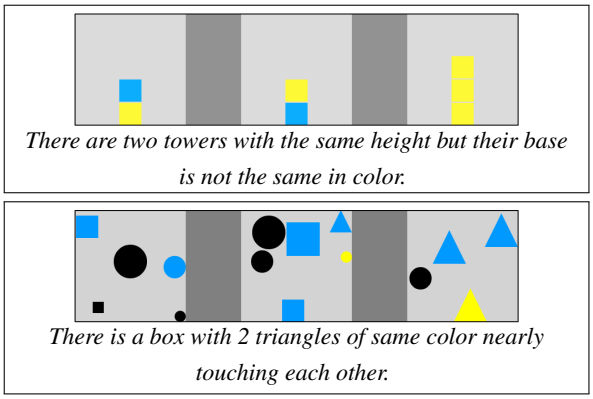


Figure 1: Example sentences and images from our corpus. Each image includes three boxes with different object types. The truth value of the top sentence is true, while the bottom is false.

1 Introduction

Understanding complex compositional language in context is a challenge shared by many tasks. Visual question answering and robot instruction systems require reasoning about sets of objects, quantities, comparisons, and spatial relations; for example, when instructing home assistance or assembly-line robots to manipulate objects in cluttered environments. This reasoning requires robust language understanding, and is only partially addressed by existing datasets. VQA (Antol et al., 2015), while lexically and visually diverse, includes relatively short sentences with limited coverage of such phenomena. CLEVR (Johnson et al., 2016) and SHAPES (Andreas et al., 2016b), in contrast, display complex compositional structure, but include only synthetic language.

In this paper, we introduce the Cornell Natural Language Visual Reasoning (NLVR) corpus and task. We define the binary prediction task of judging if a statement is true for an image or not, and introduce a corpus of annotated pairs of natural language statements and synthetic images.

Collecting this kind of language presents two challenges. First, we must design environments to support such descriptions. We use simple visual environments displaying objects with complex visual relations between them. Figure 1 shows two generated images. The second challenge is eliciting complex descriptions displaying a range of syntactic and semantic phenomena. We use a two-stage crowdsourcing process. In the first stage, we present sets of images and ask workers to write descriptive statements that distinguish them. Using synthetic images with abstract shapes allows us to control the potential distinctions between them; for example, by discouraging simple statements about object existence. In the second stage, we ask workers to label the truth value for the sentences and images generated in the first stage.

Our data includes 92,244 sentence-image pairs with 3,962 unique sentences. We include both images and the structured representation used to generate them to support research using both raw visual information and structured data. Figure 1 shows two examples. To assess the difficulty of NLVR, we experiment with multiple baselines. The best model using images achieves an accuracy of 66.12, demonstrating remaining challenges

in the data. We also analyze the language in our data for presence of certain linguistic phenomena, and compare this analysis with related datasets. The data and leaderboard are available at http://lic.nlp.cornell.edu/nlvr.

2 Related Work and Datasets

Several datasets have been created to study visual reasoning and language. VQA (Antol et al., 2015; Zitnick and Parikh, 2013) includes crowdsourced questions and answers for photographs and abstract scenes, and has been studied extensively (e.g., Lu et al., 2016; Xu and Saenko, 2016; Zhou et al., 2015; Chen et al., 2015a; Andreas et al., 2016b,a; Ray et al., 2016). In contrast to VQA, we use synthetic images and emphasize representing a broad range of language phenomena. Our motivation is similar to that of SHAPES (Andreas et al., 2016b) and CLEVR (Johnson et al., 2016). Both datasets also use synthetic images and emphasize representing diverse spatial language. However, unlike our approach, they include only automatically generated language.

Visual reasoning has also been addressed in instructional language corpora (e.g., MacMahon et al., 2006; Chen and Mooney, 2011; Bisk et al., 2016), where executable instructions are grounded in manipulable environments. The language we observe is similar to the type of language studied for understanding and generation of referential expressions (Mitchell et al., 2010; Matuszek et al., 2012; FitzGerald et al., 2013).

Our task is related to caption generation, which has been studied extensively (e.g., Pedersoli et al., 2016; Carrara et al., 2016; Chen et al., 2016) with MSCOCO (Chen et al., 2015b) and Flickr30K (Young et al., 2014; Plummer et al., 2015). In contrast to caption generation, our task does not require approximate metrics like BLEU.

Several existing datasets focus on natural language querying of structured representations, including GeoQuery (Zelle, 1995) and WikiTables (Pasupat and Liang, 2015). Our work is complementary to these resources. While our corpus was collected using images, we also provide structured representations. When used with these representations, our corpus is similar to WikiTables, where questions are paired with small web tables. Instead of web tables, we use object sets and focus on visual language.

3 Task

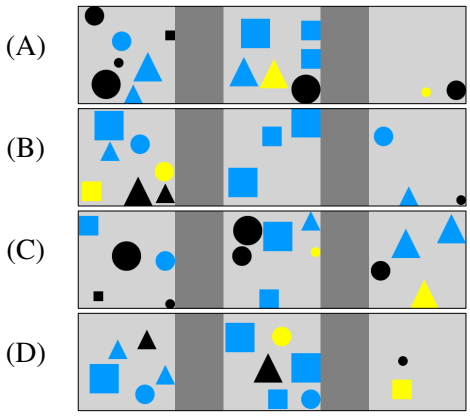
Statements in our data are grounded in synthetic images rendered from structured representations. Given an example, the task is to determine whether a statement is true or false for the image or structured representation. While we describe the image, the structured representation is equivalent. We provide examples of the structured representation in the supplementary material. Images are divided into three *boxes*. Figure 1 shows two images. Each box contains 1-8 *objects*. Each object has four properties: *position* (x/y coordinates), *color* (black, blue, yellow), *shape* (triangle, square, circle), and *size* (small, medium, large). Objects within a box cannot overlap and must be contained entirely in the box. We distinguish between images containing scattered objects and images containing only squares arranged in towers up to four blocks tall. The top image in Figure 1 is a tower example; the bottom is a scatter example.

This design encourages compositional language with complex visual reasoning. We divide the image into boxes to encourage set theoretic reasoning within and between boxes. We also use a relatively limited number of values for each property. While a large number of properties provides a more diverse image, it is likely to result in descriptions that refer to property differences. We find that the limited number of properties elicits descriptions with rich compositional structure.

4 Data Collection

We generate images following the structure described in Section 3, and collect grounded natural language descriptions. Data is collected in two phases: sentence writing and validation. During sentence writing, workers are asked to write contrasting descriptions about a set of images. To validate sentences, the description is paired with each of the images. We execute the collection process four times to collect training, development, and two test sets (Test-P and Test-U). We retain one test set as unreleased (Test-U).

Generating Images We generate images by rendering a randomly sampled structured representation. The number of objects in each box and their properties are sampled uniformly. We generate an equal number of scatter and tower images. To generate the sets of images presented to annotators, we generate two images independently, a third image by using the set of objects in the first im-



Write one sentence. This sentence must meet all of the following requirements:

- It describes A.
- It describes B.
- It does *not* describe C.
- It does *not* describe D.
- It does *not* mention the images explicitly (e.g. "In image A, ...").
- It does *not* mention the order of the light grey squares (e.g. "In the rightmost square...")

There is no one correct sentence for this image. There may be multiple sentences which satisfy the above requirements. If you can think of more than one sentence, submit only one.

Figure 2: Sentence writing prompt. The bottom sentence in Figure 1 was generated from this prompt.

age and randomly re-shuffling them between the boxes, and a fourth image by re-shuffling the objects in the second image. For images with towers, we constrain the re-shuffling to form towers.

Phase 1 – Sentence Writing Each writing task presents an annotator with four images. Figure 2 shows the sentence writing prompt, including the set of constraints, which is shown for all writing tasks. The constraints force the worker to contrast two pairs by referring to similarities and differences between the images, but not to refer to the position of the image in the prompt, or of each box in each image. These constraints are placed to elicit more set-theoretic language, and to allow us to divide the result of each task into four examples, pairing the annotator's sentence with each of the four images it was presented with.

Phase 2 – Validation In the second phase, we pair each sentence with the four images used to generate it. We re-label all sentence-image pairs as true or false, correcting for any violations of the constraints in the first phase. We do not use the original position of the image as any part of the final label to neutralize any ordering effect. In practice, 8.2% of examples had a different label than inferred from their original position in the first phase. During validation, boxes are randomly permuted to ensure the last constraint was followed. We allow workers to annotate a sentence as nonsensical with regard to the image, and instruct annotators to ignore grammar errors.

Post-processing We prune pairs when their majority class is nonsensical. When collecting multiple annotations for a pair, we prune pairs if the gap between the classes is less than two votes.

	Unique sentences	Examples
Train	3,163	74,460
Dev	267	5,940
Test-P	266	5,934
Test-U	266	5,910
Total	3,962	92,244

Table 1: Data statistics.

5 Data Statistics and Analysis

We use the crowdsourcing platform Upwork,[1] and select ten annotators using a small set of example questions. We collect 3,974 task instances and 28,723 total validation judgments at a total cost of $5,526. From these 3,974 task instances we extract 15,896 sentence-image pairs. We prune 522 pairs in post-processing. For the training set we collect a single validation annotation for each sentence-image pair; for the rest of the data we collect five annotations each. Finally, we generate six sentence-image pairs from each sample by permuting the boxes. The validation step ensures this permutation does not change the label. Table 1 shows the number of sentences and pairs, including permutations, for each split.

We merge the development and test splits to calculate agreement statistics. We calculate Krippendorf's α and Fleiss' κ (Cocos et al., 2015) on both the full and pruned datasets. To calculate Fleiss' κ, we randomly permute the five annotations to be assigned to five "raters" and compute average kappa from 100 iterations. Before pruning, we observe $\alpha = 0.768$ and $\kappa = 0.709$, indicating substantial agreement (Landis and Koch, 1977). Pruning improves agreement to $\alpha = 0.831$ (indicating almost-perfect agreement) and $\kappa = 0.808$.

We analyze 200 development sentences to identify the distribution of semantic phenomena and syntactic ambiguity (Table 2). For comparison, we apply this analysis to 200 abstract-image and 200 real-image sentences from VQA (Antol et al., 2015). The difference in the distribution illustrates the complexity of our data. The mean sentence

[1] http://upwork.com

	VQA (abs)	VQA (real)	Our Data	NMN Correct	Example
Semantics					
Cardinality (hard)	12	11.5	66	63.8	*There are **exactly four objects** not touching any edge*
Cardinality (soft)	0	1	16	63.4	*There is a box with **at least one** square and **at least three** triangles.*
Existential	4.5	11.5	88	64.2	***There is a tower** with yellow base.*
Universal	1	1	7.5	67.8	*There is a black item in **every box.***
Coordination	3	5	17	58.5	*There are 2 blue circles **and** 1 blue triangle*
Coreference	8.5	6.5	3	55.3	*There is a blue triangle touching the wall with **its** side.*
Spatial Relations	31	42.5	66	61.6	*there is one tower with a yellow block **above** a yellow block*
Comparative	1.5	1	3	73.6	*There is a box with multiple items and only one item **has a different color**.*
Presupposition[2]	79	80	19.5	54.0	*There is a box with seven items and **the three black items** are the same in shape.*
Negation	0	1	9.5	51.0	*there is exactly one black triangle **not touching** the edge*
Syntax					
Coordination	0	0	4.5	53.4	*There is a box with at least one square **and** at least three triangles.*
PP Attachment	7	3	23	70.9	*There is a black block on a black block as the base of a tower **with** three blocks.*

Table 2: Qualitative and empirical analysis of our data and VQA (Antol et al., 2015). We analyze 200 sentences for each dataset. The data is categorized to semantic and syntactic categories. We use the terms *hard* and *soft* cardinality to differentiate between language using exact numerical values and ranges. For each dataset, we show the percentage of the samples analyzed that demonstrate the phenomena. We analyze abstract (abs) and real images from VQA separately. For our data, we also include the accuracy using the NMN system (Section 6) for the subset of images we tagged with this category.

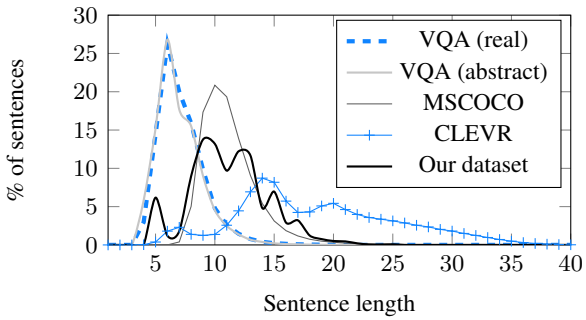


Figure 3: Distribution of sentence lengths.

length in our data is 11.22 tokens and the vocabulary size is 262. In Figure 3, we compare sentence length distribution to VQA, MSCOCO (Chen et al., 2015b), and CLEVR (Johnson et al., 2016). Our sentences are generally longer than VQA and more similar in length to MSCOCO. However, our task is more similar to VQA, where context is used to understand language, rather than to generate.

6 Methods

We evaluate multiple methods on the rendered images and structured representations. Hyperparameters and initialization details are described in the supplementary material.

[2]We say a statement or question uses presupposition when it assumes the truth value of some proposition in order for its entire truth value to be defined. In this example, an image which does not have three black items will have no defined truth value for this statement.

6.1 Majority Class and Single Modality

We use image- and text-only models to measure how well biases in our data can be used to solve the task. If the model is able to do well on the text- or image-only baselines, this implies our data does not require the two modalities. Antol et al. (2015) performed a similar analysis of VQA with the questions only to gauge how and if background knowledge of the domain could aid performance.

Majority Assign the most common label (`true`) to all examples.

Text Only Encode the sentence with a recurrent neural network (RNN; Elman, 1990) with long short-term memory units (LSTM; Hochreiter and Schmidhuber, 1997) and a binary softmax computed from the final output.

Image Only Encode the image with a convolutional neural network (CNN) with three layers. The CNN output is used by a three-layer perceptron with a softmax on the final layer.[3]

6.2 Structured Representation

We use the structured representations described in Sections 3 and 4.

[3]We also experimented using the ImageNet-trained Inception v4 model (Szegedy et al., 2017), but found it did not improve performance, possibly due to the difference between our images and ImageNet.

		Train	Dev	Test-P	Test-U
	Majority	56.37	55.31	56.16	55.43
	Text only	58.36 ±0.6	56.61 ±0.5	57.18 ±0.6	56.21 ±0.4
	Image Only	56.79 ±1.3	55.35 ±0.1	56.05 ±0.3	55.33 ±0.3
Structured representation	MaxEnt	99.99	68.04	67.68	67.82
	MLP	96.15 ±1.3	67.50 ±0.5	66.28 ±0.4	65.32 ±0.4
	Image features+RNN	59.71 ±1.0	57.72 ±1.4	57.62 ±1.3	56.29 ±0.9
Raw image	CNN+RNN	58.85 ±0.2	56.59 ±0.3	58.01 ±0.3	56.30 ±0.6
	NMN	98.37±0.6	63.06 ±0.1	66.12 ±0.4	61.99 ±0.8

Table 3: Mean accuracy and standard deviation results. We report accuracy for the train, development, and both test sets. Three systems use the structured representation. Two systems (and Image Only) use the raw image.

MaxEnt Train a MaxEnt classifier. We use the text and structured representation to compute property- and count-based features. Property-based features trigger when some property (e.g., an object is touching a wall) is true in the structure. We create features by crossing triggered properties with each n-grams from the sentence, up to $n = 6$. Count-based features trigger when a count we observe in the image (e.g., the number of black triangles) is present in the sentence. We generate features combining the type of item counted (e.g., black triangles) with the n-grams surrounding the count in the sentence, up to $n = 6$. We provide details in the supplementary material.

MLP Train a single-layer perceptron with a softmax layer. The input to the perceptron is the mean of the feature embeddings. We use the same feature set as the MaxEnt model.

Image Features+RNN Compute features from the structure representation only, and encode the text with an LSTM RNN. The two representations are concatenated, and used as input to a two-layer perceptron and a softmax layer.

6.3 Image Representation

CNN+RNN Concatenate the CNN and RNN representations (Section 6.1) and apply a multi-layer perceptron with a softmax.

NMN The neural module networks approach of Andreas et al. (2016b). We experiment with the default maximum leaves of two, and with allowing for more expressive representations with a maximum leaves of five. We observe higher development accuracy with the trees using maximum leaves of five (63.06% vs. 62.4% with the default of two), which we use in our experiments.

7 Results

We run each experiment ten times and report mean accuracy as well as standard deviation for randomly initialized models. Table 3 shows our results. NMN is the best performing model using images. Table 2 shows the NMN accuracy for each category in our qualitative analysis sample. While the number of sentences in some categories is relatively small, we observe a higher number of failures in sentences that include negations and coordinations. For models using the structured representation, the MaxEnt model provides the best performance. When ablating count-based features from the MaxEnt model, development accuracy decreases from 68.04 to 57.7. This indicates counting is an important aspect of the problem.

8 Discussion

We introduce the Cornell Natural Language Visual Reasoning dataset and task. The data includes complex compositional language grounded in images and structured representations. The task requires addressing challenges in visual and set-theoretic reasoning. We experiment with multiple systems and, in general, observe relatively low performance. Together with our qualitative analysis, this exemplifies the complexity of the data. We release our annotated training and development sets, and create two test sets. The public test set will be released along with its annotation. Computing results on the unreleased test data will require submitting trained models. Procedures for submitting models and the task leader board are available at http://lic.nlp.cornell.edu/nlvr.

Acknowledgments

This research was supported by a Microsoft Research Women's Fellowship, a Google Faculty Award, and an Amazon Web Services Cloud Credits for Research Grant. We thank the Cornell and University of Washington NLP groups for their support and helpful comments. We thank the anonymous reviewers for their feedback.

References

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016a. Learning to compose neural networks for question answering. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. https://doi.org/10.18653/v1/N16-1181.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016b. Neural module networks. In *Conference on Computer Vision and Pattern Recognition*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *International Journal of Computer Vision*.

Yonatan Bisk, Deniz Yuret, and Daniel Marcu. 2016. Natural language communication with robots. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. https://doi.org/10.18653/v1/N16-1089.

Fabio Carrara, Andrea Esuli, Tiziano Fagni, Fabrizio Falchi, and Alejandro Moreo. 2016. Picture it in your mind: Generating high level visual representations from textual descriptions. *CoRR* abs/1606.07287.

David L. Chen and Raymond J. Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the National Conference on Artificial Intelligence*.

Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ramakant Nevatia. 2015a. ABC-CNN: An attention based convolutional neural network for visual question answering. *CoRR* abs/1511.05960.

Wenhu Chen, Aurélien Lucchi, and Thomas Hofmann. 2016. Bootstrap, review, decode: Using out-of-domain textual data to improve image captioning. *CoRR* abs/1611.05321.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015b. Microsoft COCO captions: Data collection and evaluation server. *CoRR* abs/1504.00325.

Anne Cocos, Aaron Masino, Ting Qian, Ellie Pavlick, and Chris Callison-Burch. 2015. Effectively crowdsourcing radiology report annotations. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*. https://doi.org/10.18653/v1/W15-2614.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science* 14:179–211.

Nicholas FitzGerald, Yoav Artzi, and Luke Zettlemoyer. 2013. Learning distributions over logical forms for referring expression generation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. http://www.aclweb.org/anthology/D13-1197.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9:1735–1780.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2016. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *CoRR* abs/1612.06890.

J. Richard Landis and Gary Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33 1:159–74.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Neural Information Processing Systems*.

Matthew MacMahon, Brian Stankiewics, and Benjamin Kuipers. 2006. Walk the talk: Connecting language, knowledge, action in route instructions. In *Proceedings of the National Conference on Artificial Intelligence*.

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A joint model of language and perception for grounded attribute learning. In *Proceedings of the International Conference on Machine Learning*.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2010. Natural reference to objects in a visual domain. In *Proceedings of the 6th International Natural Language Generation Conference*. http://aclweb.org/anthology/W10-4210.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. https://doi.org/10.3115/v1/P15-1142.

Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. 2016. Areas of attention for image captioning. *CoRR* abs/1612.01033.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *The IEEE International Conference on Computer Vision*.

Arijit Ray, Gordon Christie, Mohit Bansal, Dhruv Batra, and Devi Parikh. 2016. Question relevance in VQA: Identifying non-visual and false-premise

questions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. http://aclweb.org/anthology/D16-1090.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Association for the Advancement of Artificial Intelligence*.

Huijuan Xu and Kate Saenko. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations. *Transactions of the Association of Computational Linguistics* http://aclweb.org/anthology/Q14-1006.

John M. Zelle. 1995. *Using inductive logic programming to automate the construction of natural language parsers*. Ph.D. thesis, University of Texas, Austin.

Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2015. Simple baseline for visual question answering. *CoRR* abs/1512.02167.

C. Lawrence Zitnick and Devi Parikh. 2013. Bringing semantics into focus using visual abstraction. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*.