Supplementary Material: Natural Language Object Retrieval

Ronghang Hu[1] Huazhe Xu[2] Marcus Rohrbach[1,3] Jiashi Feng[4] Kate Saenko[5] Trevor Darrell[1]
[1]University of California, Berkeley [2]Tsinghua University [3]ICSI, Berkeley
[4]National University of Singapore [5]University of Massachusetts, Lowell
{ronghang, rohrbach, trevor}@eecs.berkeley.edu, xhz12@mails.tsinghua.edu.cn
elefjia@nus.edu.sg, saenko@cs.uml.edu

Abstract

In this document, we visualize some results on the ReferIt dataset [1] using our SCRC model, showing that it can correctly retrieve an object by exploiting its description in context. We also evaluate our model on the Flickr30K Entities dataset [2], and show that our model can be applied to both "object" and "stuff", and can generate descriptions over given image regions.

1. Retrieval on object descriptions in context

In reality, people usually describe an object based on both the object itself and other objects plus the whole scene as context. To distinguish a specific object from others in a scene, especially when there are multiple objects of the same category, a description needs to contain not only the category name, but also other discriminative information such as locations or attributes.

Figure 1 shows an example of this, where one cannot refer to a person simply using category name "person" since there are three people in the scene, but needs to use a description based on the environment as query. Our SCRC model can handle such context-based descriptions by incorporating spatial configurations and scene-level context into the recurrent network. Figure 2 shows some retrieval examples on multiple objects within the same image on ReferIt [1] dataset, where objects are described in context.

2. Object retrieval evaluation on Flickr30K Entities dataset

We also train and evaluate our method on the Flickr30K Entities dataset [2] for natural language object retrieval, which contains 31,783 images and 275,775 annotated bounding boxes. The object-level annotations in this dataset are derived from existing scene-level captions in Flickr30K [3].

We train our model on the referential expressions in the Flickr30K dataset using the same top-100 EdgeBox [4] proposals same as in [2]. On this dataset, our SCRC model achieves higher recall than the Canonical Correlation Analysis (CCA) method in [2], as is shown in Table 1.

Method	R@1	R@10
CCA [2]	25.3%	59.7%
SCRC	**27.8%**	**62.9%**
Oracle	76.9%	76.9%

Table 1. Performance of our method compared with Canonical Correlation Analysis (CCA) baseline on 100 EdgeBox proposals in Flickr30K Entities dataset. Oracle corresponds to the highest possible recall on all 100 proposals for any retrieval method.

3. Object vs. stuff

The ReferIt dataset contains annotations on both "object" regions and "stuff" regions. In computer vision, the term **object** is usually used to refer to entities with closed boundary and well-defined shape, such as "car", "person" and "laptop". On the other hand, **stuff** is used for entities without a regular shape, such as "grass", "road" and "sky".

Given an input image and a natural language query, our SCRC model is not only capable of retrieving "object" regions, but can also be applied to "stuff" regions. Figure 3 shows some examples of stuff retrieval on ReferIt dataset.

4. Generating descriptions for objects

Although our SCRC model is designed for natural language object retrieval, it can also be applied in another task to generate descriptions for the objects in an image. Given an image I_{im} and the bounding box of an object, a text description S_{des} can be generated for the object as $S_{des} = \arg\max_S p(S|I_{box}, I_{im}, x_{spatial})$ using beam search, where I_{im} is the local image region of the object and $x_{spatial}$ is its spatial configuration. Figure 4 shows some object descriptions generated by our SCRC model on ReferIt dataset.

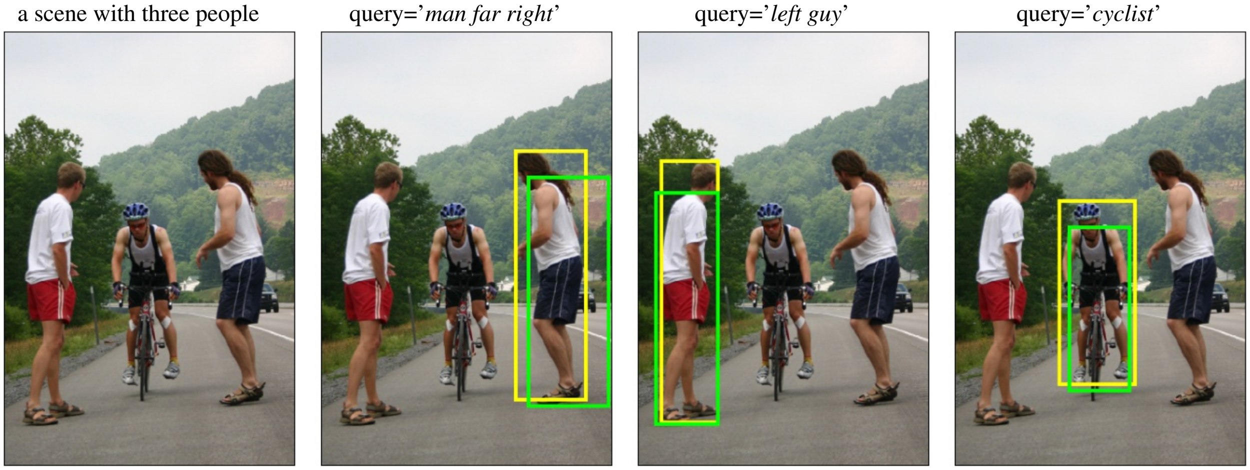


Figure 1. An example image in ReferIt dataset where objects are described based on other objects in the scene. When referring to one of the three "people" in the image, expressions based on both the object and the context are used to make the description discriminative. Our model can handle such object descriptions in context by incorporating these information into the recurrent neural network. In the images above, yellow boxes are ground truth and green boxes are correctly retrieved results by our model using highest scoring candidate from 100 EdgeBox proposals.

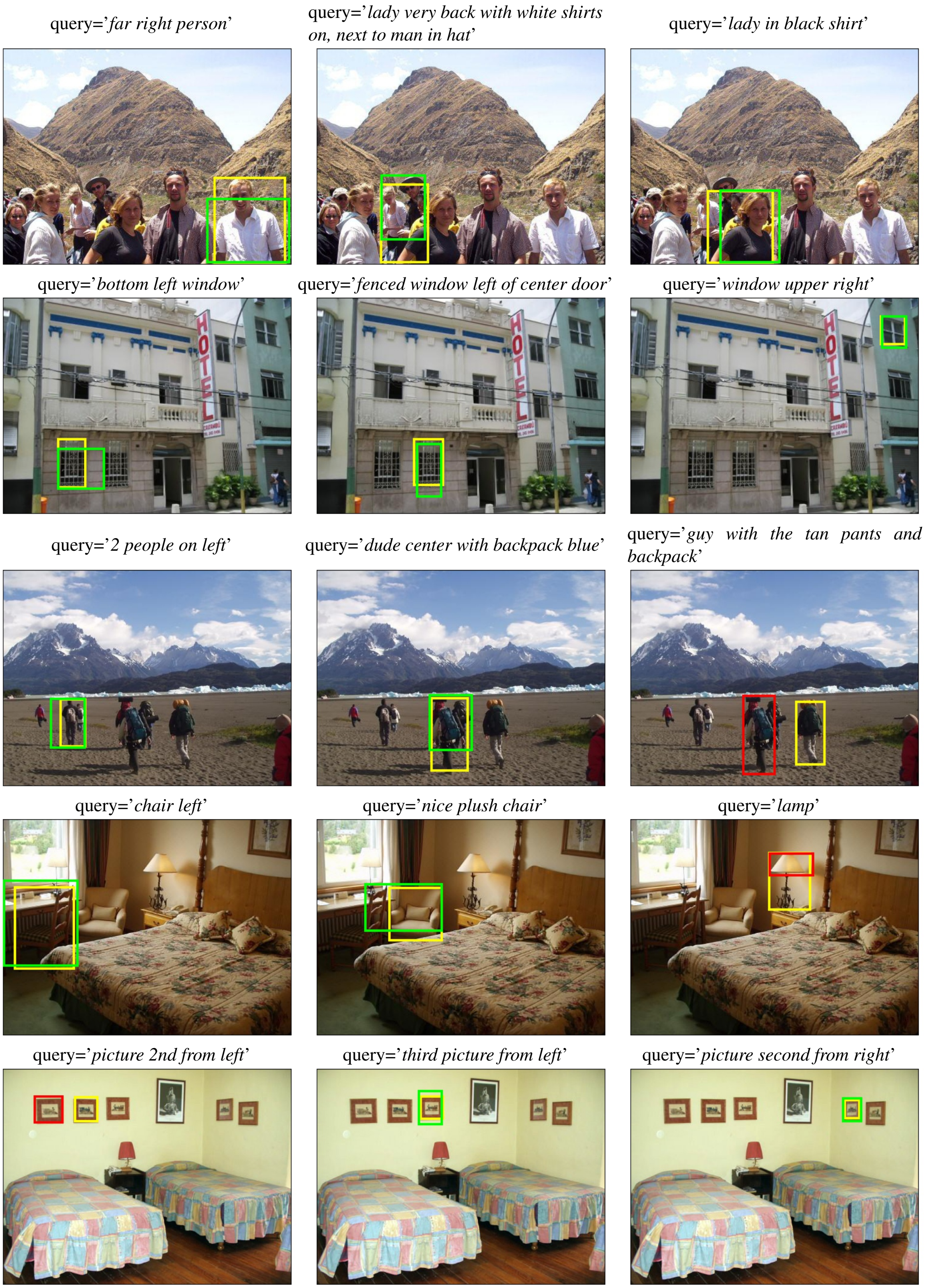
References

[1] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, 2014. 1

[2] B. Plummer, L. Wang, C. Cervantes, J. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 1

[3] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 1

[4] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 391–405. Springer, 2014. 1



query='*far right person*'

query='*lady very back with white shirts on, next to man in hat*'

query='*lady in black shirt*'

query='*bottom left window*'

query='*fenced window left of center door*'

query='*window upper right*'

query='*2 people on left*'

query='*dude center with backpack blue*'

query='*guy with the tan pants and backpack*'

query='*chair left*'

query='*nice plush chair*'

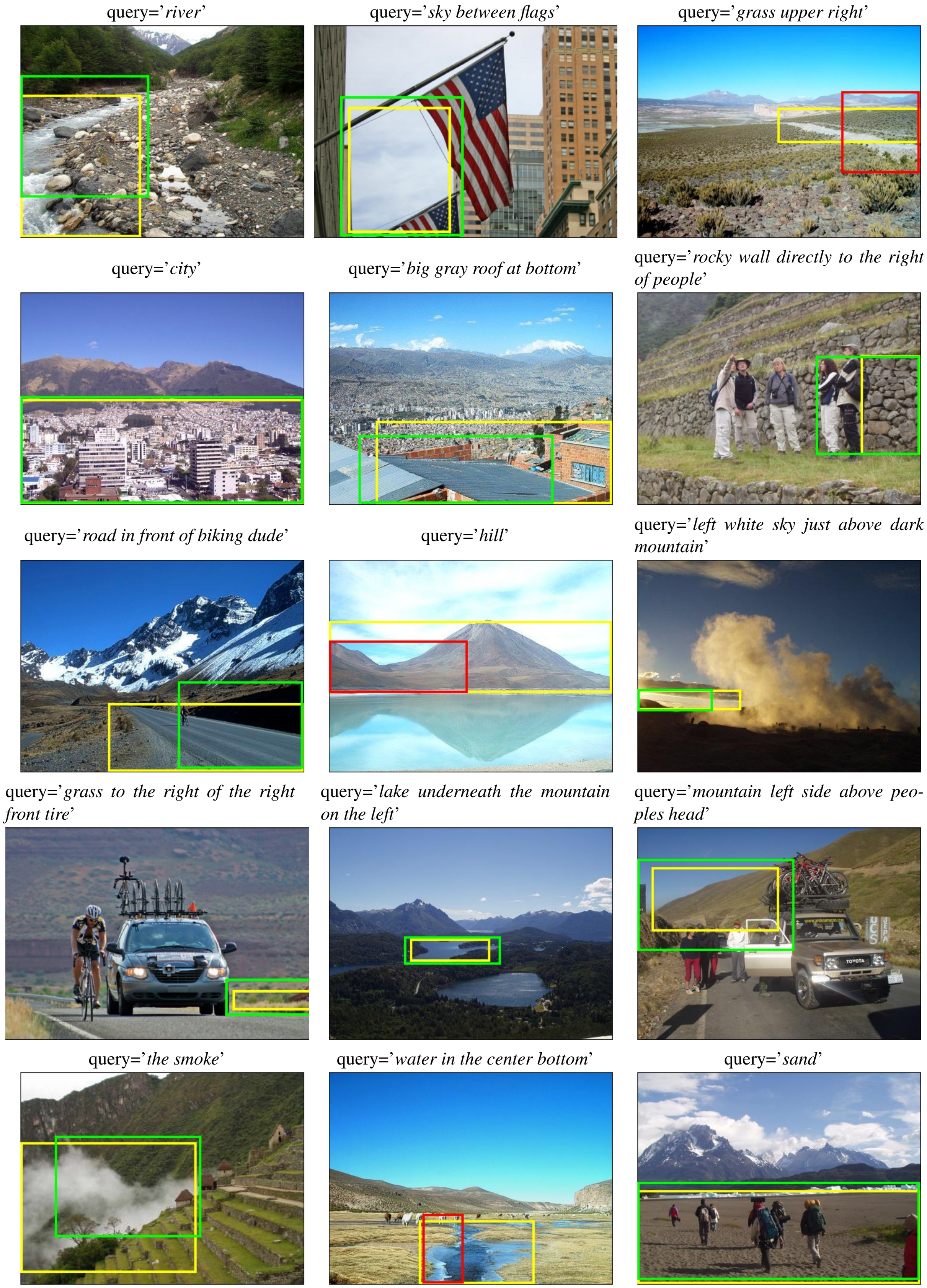
query='*lamp*'

query='*picture 2nd from left*'

query='*third picture from left*'

query='*picture second from right*'

Figure 2. Examples on multiple objects in the same image in ReferIt, showing the highest scoring candidate box (correct in green, incorrect in red) from 100 EdgeBox proposals and ground truth (yellow). Our model retrieves the objects by taking their local descriptors, spatial configurations and scene-level contextual information into account.



query='*river*'

query='*sky between flags*'

query='*grass upper right*'

query='*city*'

query='*big gray roof at bottom*'

query='*rocky wall directly to the right of people*'

query='*road in front of biking dude*'

query='*hill*'

query='*left white sky just above dark mountain*'

query='*grass to the right of the right front tire*'

query='*lake underneath the mountain on the left*'

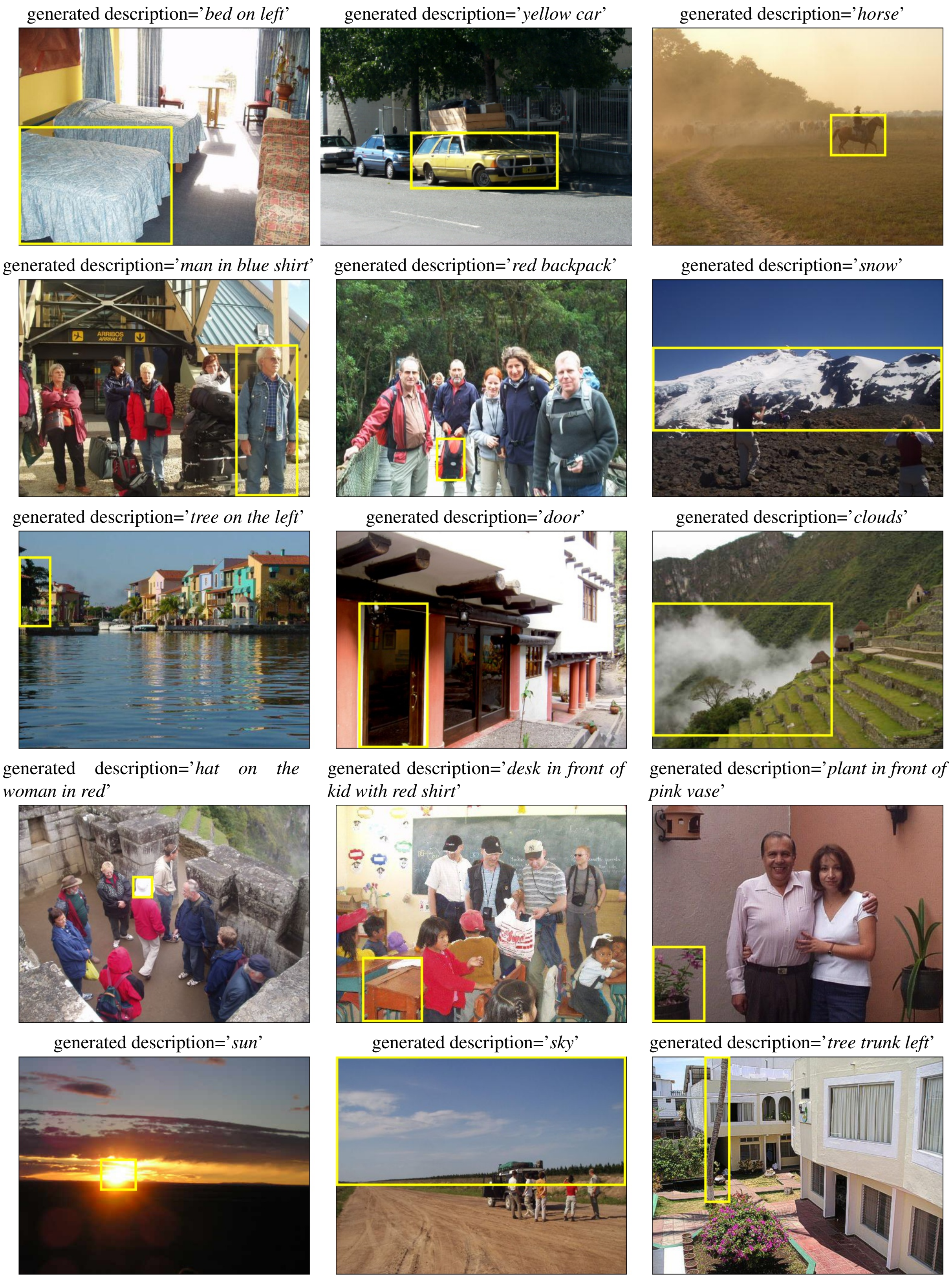
query='*mountain left side above peoples head*'

query='*the smoke*'

query='*water in the center bottom*'

query='*sand*'

Figure 3. Examples on "stuff" regions in ReferIt, showing the highest scoring candidate box (correct in green, incorrect in red) from 100 EdgeBox proposals and ground truth (yellow).



generated description=’*bed on left*’

generated description=’*yellow car*’

generated description=’*horse*’

generated description=’*man in blue shirt*’

generated description=’*red backpack*’

generated description=’*snow*’

generated description=’*tree on the left*’

generated description=’*door*’

generated description=’*clouds*’

generated description=’*hat on the woman in red*’

generated description=’*desk in front of kid with red shirt*’

generated description=’*plant in front of pink vase*’

generated description=’*sun*’

generated description=’*sky*’

generated description=’*tree trunk left*’

Figure 4. Generated object descriptions by our model on ReferIt. The bounding box of the object is shown in yellow.