Supplementary Material for "Cops-Ref: A new Dataset and Task on Compositional Referring Expression Comprehension"

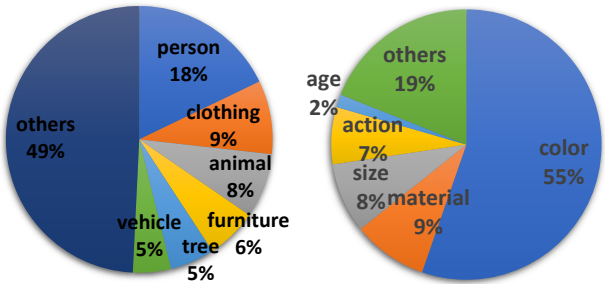
Zhenfang Chen[1*] Peng Wang[2] Lin Ma[3] Kwan-Yee K. Wong[1] Qi Wu[4†]
[1]The University of Hong Kong [2]University of Wollongong
[3]Tencent AI Lab [4]Australian Centre for Robotic Vision, University of Adelaide
[1]{zfchen, kykwong}@cs.hku.hk [2]pengw@uow.edu.au
[3]forest.linma@gmail.com [4]qi.wu01@adelaide.edu.au

This document aims to provide additional materials to supplement our main submission. We first show more statistics in Sec. 1. Then, we give more details on the implementation of the baseline method in Sec. 2. Finally, we show more experimental results in Sec. 3, including examples of hard mining training strategy, qualitative results and a failure case.

1. Statistics on Cops-Ref

We merge the entry-level object categories and attributes into higher-level categories (*e.g.* we merge "skirt", "short", "t-shirt" and *etc.* into the high-level category "clothing"). We show the distributions of the high-level object categories and attribute categories in Fig. 1. From Fig. 1 (a), we find that the dataset has a diverse range of object classes, where person, clothing, animal and furniture are the most popular classes and others covering 49% indicates other small object classes whose proportion is less than 5%. In Fig. 1 (b), we summarise of the distributions of the attribute categories, where simple attributes like color, material and sizes are the most frequent. In fact, such simple attributes reflect the natural tendencies of the real world because people usually tend to use simple and distinguishable attributes to describe objects.

We show more examples of the proposed Cops-Ref dataset in Fig. 2-7. From these examples, we can observe two main features of the Cops-Ref dataset. First, the expressions of the dataset are flowery and compositional with different reasoning logics represented by the corresponding reasoning trees. Second, the dataset contains diverse distracting images with varying distracting factors such as objects of the same category, the same attributes and object interactions. Moreover, these two features are complementary. The flowery and compositional expressions make it possible to distinguish the target object from the similar instances in the distracting images. On the other hand, the semantically similar distractors guarantee that good performance can only be achieved by a model that fully understands the complete expressions.



(a) Object categories (b) attribute categories.

Figure 1: The distribution of object categories and attribute categories

2. Implementation Details

For GroundeR [5], it uses a global feature to represent each object proposal. We get the proposal feature by concatenating the averaged-pooled object proposal features from the C3 and C4 layers of the res101-based Faster-RCNN [1, 4]. Since it has been proved in [7, 3] that the position embeddings for the bounding boxes are good for referring expression comprehension, we additionally concatenate the coordinates of the bounding boxes into the concatenated feature. For the text-based object retrieval model SCAN [2], we use the same global concatenated feature to represent each object proposal.

3. Experiments

Mining Example We show a typical hard mining example of the mining strategy in Fig. 8. We show the positive

*Work done while Zhenfang Chen was visiting the University of Adelaide.

†Corresponding author.

region-expression pair and their corresponding modular attentive weights for each word in Fig 8 (a). We can see that the language attention network can attend to the right words for each module. We show the most similar expressions for each module and their corresponding visual regions in Fig 8 (b)-(d). We observe that the modular hard mining strategy can automatically find the corresponding hard examples. Specifically, the expression mined by the *sub* module has the same attribute "young" as the positive expression; the expression mined by the *loc* module has the same relation "to the left of"; the expression mined by the *cxt* module has the same context "white shirt".

Qualitative Examples We show two qualitative examples that the proposed MatNet-Mine outperforms the bsaeline method MattNet [6] and CM-Att-Erase [3] in Fig. 9. Since there are around 20 object proposals for each image, it will be difficult and unnecessary for us to visualise all of them in the examples. Instead, we only keep the proposal with the highest matching score for each image. We sort 13 most similar objects by their matching scores and show their ranks below the images (smaller rank means larger similarity and rank 1 denotes the prediction of the model). From the results of Fig. 9, we can see that MattNet [6] and CM-Att-Erase [3] may be confused by the distracting regions with the same categories and attributes and similar image context. They provides higher matching scores (smaller ranks) for those distracting regions than the target region. On the other hand, our model can distinguish such distracting regions to some extent because we have mined the semantically similar expressions and distracting regions during the training.

Failure Cases We also visualise a failure case in Fig. 10 in the same way as the qualitative examples in Fig. 9. From Fig. 10, we observe that all the models provide higher matching scores (small ranks) for the regions in Fig. 10 (e) than the region in Fig. 10 (a). We think the reasons are the fist and second images in Fig. 10 (e) provide strong distractions for the models. They match most parts of the expressions well and the only difference is the "pillow" is not "under" the dog. This indicates that the performance can be further improved by REF models with stronger reasoning ability for understanding the whole expression and distinguishing the subtle visual differences.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[2] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 201–216, 2018. 1

[3] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *CVPR*, 2019. 1, 2, 6, 7

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1

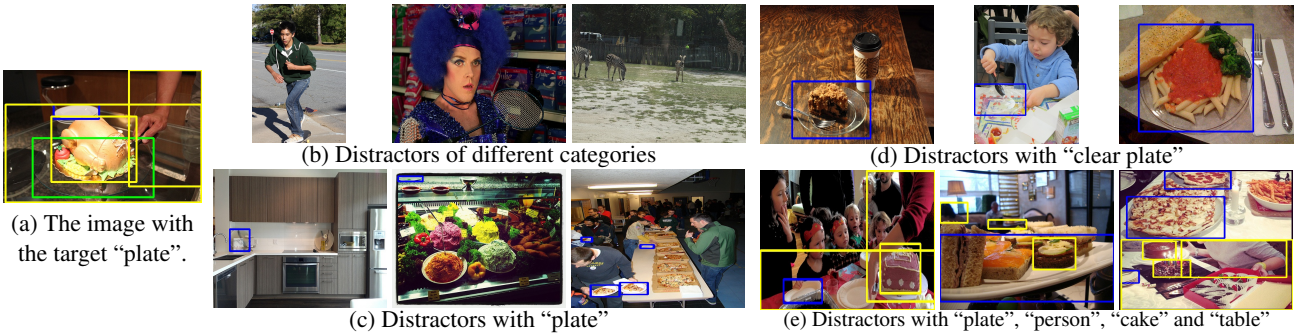
[5] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016. 1

[6] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, 2018. 2, 6, 7

[7] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 1

Reasoning tree: plate (clear) $\xrightarrow{\text{to the left of}}$ person $\xrightarrow{\text{cutting}}$ case (large) $\xrightarrow{\text{on}}$ table (clear)

Expression: *The clear plate that is to the left of the person that is cutting the large cake that is on the clear table.*



(a) The image with the target "plate".

(b) Distractors of different categories

(c) Distractors with "plate"

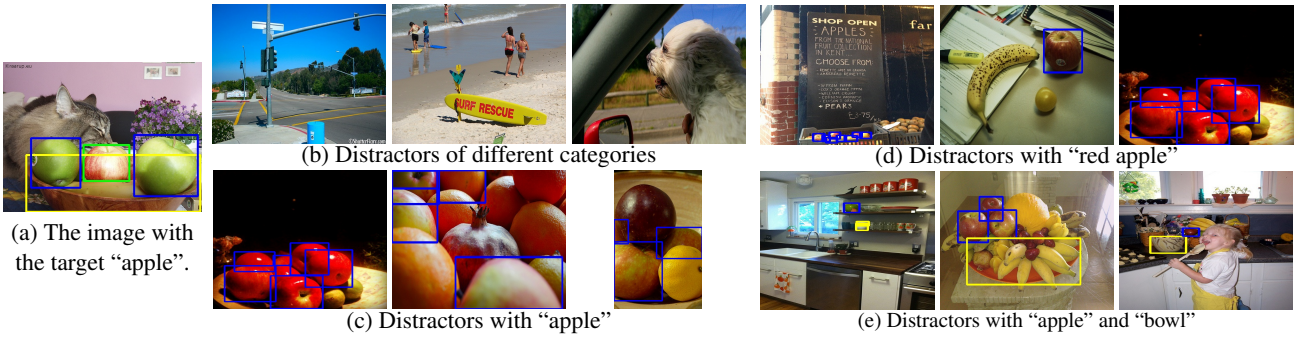
(d) Distractors with "clear plate"

(e) Distractors with "plate", "person", "cake" and "table"

Figure 2: Examples of the proposed Cops-Ref dataset. The target/related/distracting regions are marked by green/yellow/blue boxes, respectively. Related regions mean objects appear in the reasoning tree but is of different category as the target object. Distracting regions denote other objects of the same category as the target one. Best viewed on screen.

Reasoning tree: Apple (middle, left) $\xrightarrow{\text{in}}$ bowl (wood)

Expression: *Apple in the middle that is red and in wood bowl.*



(a) The image with the target "apple".

(b) Distractors of different categories

(c) Distractors with "apple"

(d) Distractors with "red apple"

(e) Distractors with "apple" and "bowl"

Figure 3: Examples of the proposed Cops-Ref dataset. The target/related/distracting regions are marked by green/yellow/blue boxes, respectively. Related regions mean objects appear in the reasoning tree but is of different category as the target object. Distracting regions denote other objects of the same category as the target one. Best viewed on screen.

Reasoning tree: couch (gray) $\begin{cases} \xrightarrow{\text{to the left of}} \text{chair (wood)} \\ \xrightarrow{\text{to the right of}} \text{chair (green)} \end{cases}$

Expression: *The gray couch that is to the left of the wood chair or to the right of the green chair.*

(a) The image with the target "couch".

(b) Distractors of different categories

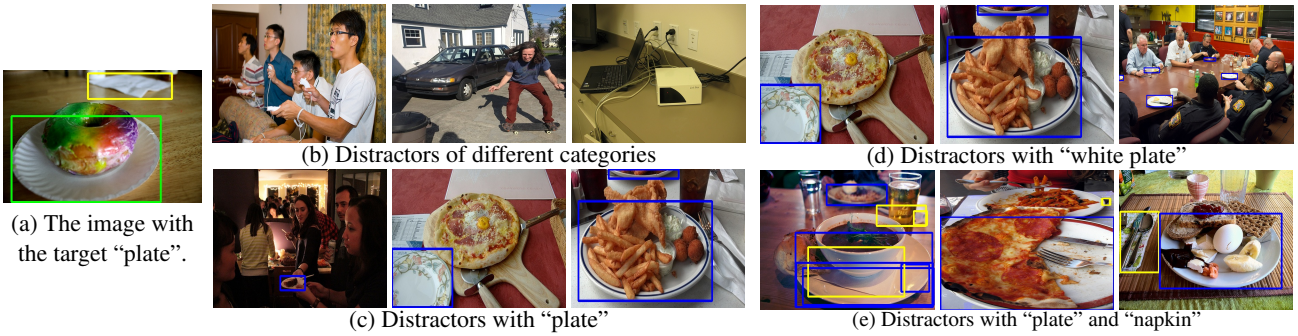
(c) Distractors with "couch"

(d) Distractors with "gray couch"

(e) Distractors with "couch" and "chair"

Figure 4: Examples of the proposed Cops-Ref dataset. The target/related/distracting regions are marked by green/yellow/blue boxes, respectively. Related regions mean objects appear in the reasoning tree but is of different category as the target object. Distracting regions denote other objects of the same category as the target one. Best viewed on screen.

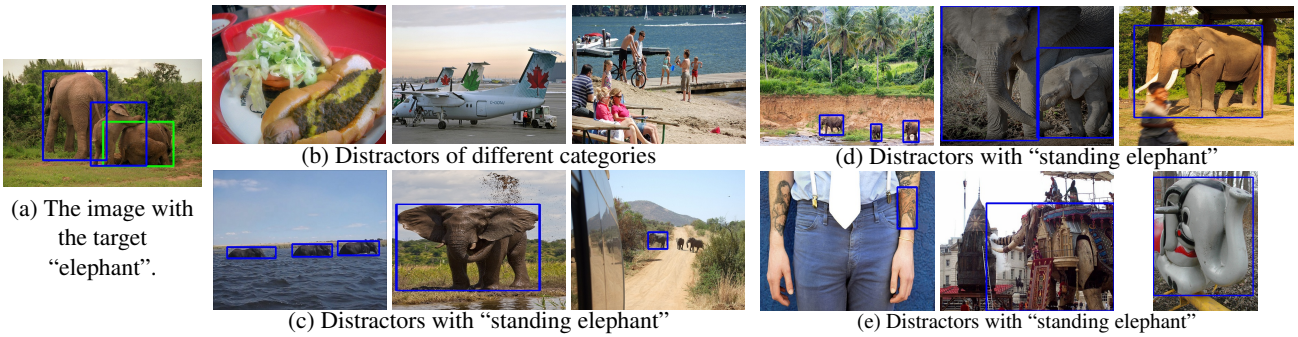
Reasoning tree: plate (white, paper) $\xrightarrow{\text{same color and material}}$ napkin (white, paper)
Expression: *The plate that has the same color and material as the napkin.*



(a) The image with the target "plate".
(b) Distractors of different categories
(c) Distractors with "plate"
(d) Distractors with "white plate"
(e) Distractors with "plate" and "napkin"

Figure 5: Examples of the proposed Cops-Ref dataset. The target/related/distracting regions are marked by green/yellow/blue boxes, respectively. Related regions mean objects appear in the reasoning tree but is of different category as the target object. Distracting regions denote other objects of the same category as the target one. Best viewed on screen.

Reasoning tree: elephant (not standing)
Expression: *The elephant that is not standing.*



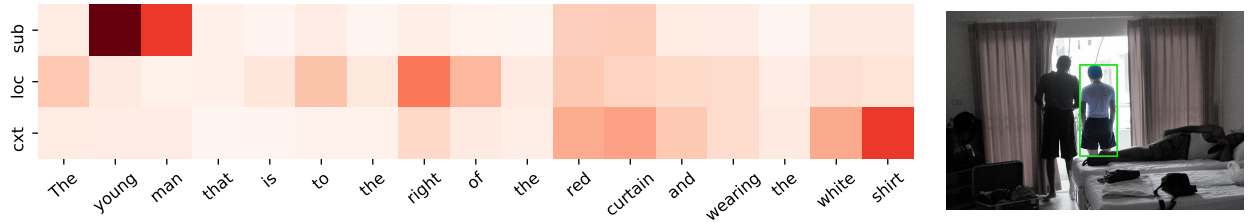
(a) The image with the target "elephant".
(b) Distractors of different categories
(c) Distractors with "standing elephant"
(d) Distractors with "standing elephant"
(e) Distractors with "standing elephant"

Figure 6: Examples of the proposed Cops-Ref dataset. The target/related/distracting regions are marked by green/yellow/blue boxes, respectively. Related regions mean objects appear in the reasoning tree but is of different category as the target object. Distracting regions denote other objects of the same category as the target one. Best viewed on screen.

Reasoning tree: woman (blond) $\begin{cases} \xrightarrow{\text{on}} \text{chair (brown)} \\ \xrightarrow{\text{using}} \text{laptop (open)} \end{cases}$
Expression: *The blond woman that is on the brown chair and using the open laptop.*

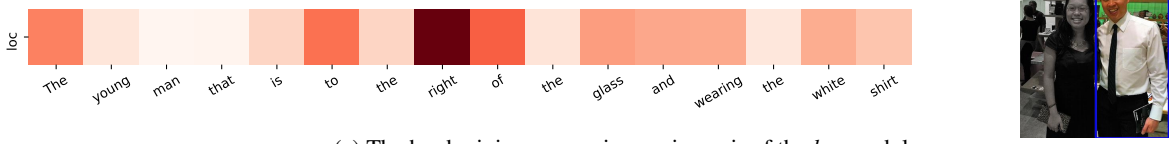
(a) The image with the target "woman".
(b) Distractors of different categories
(c) Distractors with "woman"
(d) Distractors with "blond woman"
(e) Distractors with "woman", "chair" and "laptop"

Figure 7: Examples of the proposed Cops-Ref dataset. The target/related/distracting regions are marked by green/yellow/blue boxes, respectively. Related regions mean objects appear in the reasoning tree but is of different category as the target object. Distracting regions denote other objects of the same category as the target one. Best viewed on screen.

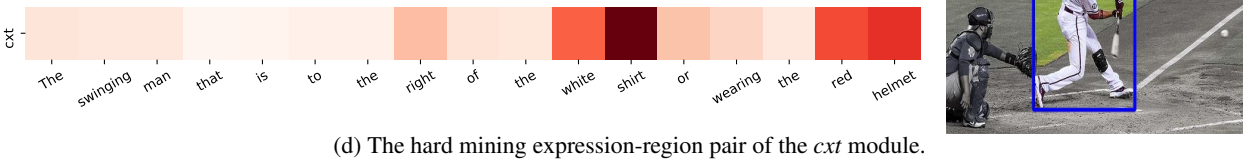


(a) The positive expression-region pair and their module attentive weight.

(b) The hard mining expression-region pair of the *sub* module.



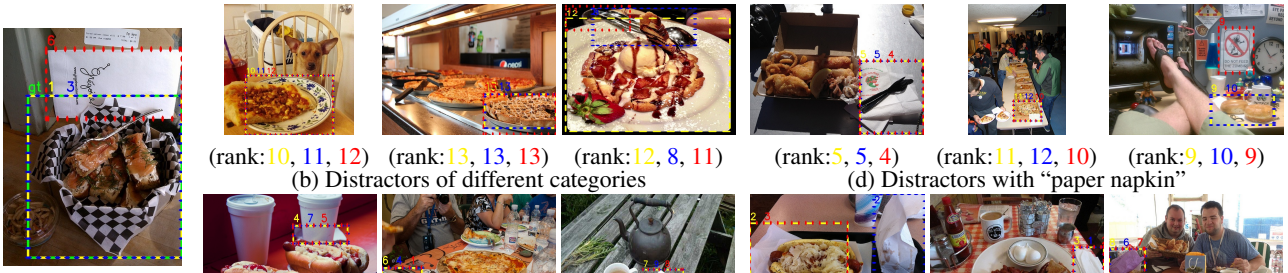
(c) The hard mining expression-region pair of the *loc* module.



(d) The hard mining expression-region pair of the *cxt* module.

Figure 8: An mining example of modular hard mining strategy. We show the most similar expressions of each module and their corresponding regions. The corresponding distributions of the attentive weights for each module are represented by colours and darker colours mean larger attentive weights. We can see that the modular harding strategy can automatically find hard mining examples for each module. The expression mined by the *sub* module has the same attribute "young" as the query expression; the expression mined by the *loc* module has the same relation "to the right of"; the expression mined by the *cxt* module has the same context "white shirt".

Expression: *The paper napkin that is in the lined basket that is near the plastic cup.*



(rank:1, 3, 6)
(a) The image with the target "napkin".

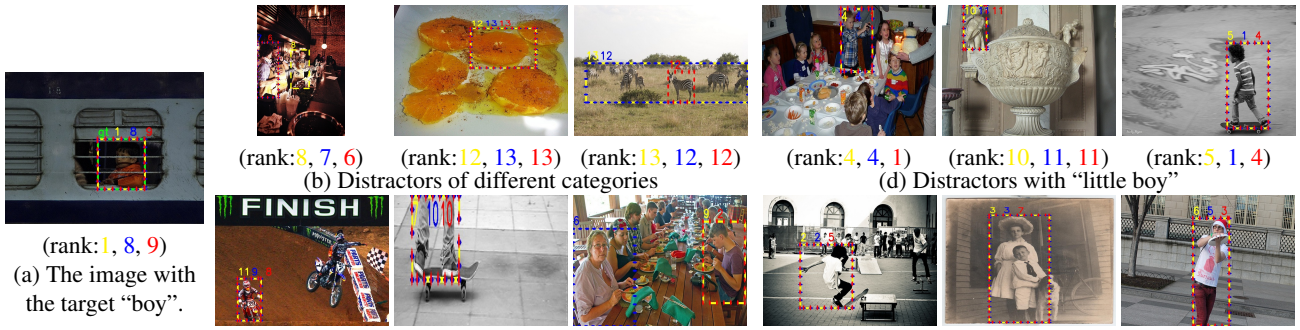
(rank:10, 11, 12) (rank:13, 13, 13) (rank:12, 8, 11)
(b) Distractors of different categories

(rank:5, 5, 4) (rank:11, 12, 10) (rank:9, 10, 9)
(d) Distractors with "paper napkin"

(rank:4, 7, 5) (rank:6, 4, 1) (rank:7, 9, 8)
(c) Distractors with "napkin"

(rank:2, 2, 3) (rank:3, 1, 2) (rank:8, 6, 7)
(e) Distractors with "napkin", "basket" and "cup"

Expression: *The little boy that is looking out the steel window.*



(rank:1, 8, 9)
(a) The image with the target "boy".

(rank:8, 7, 6) (rank:12, 13, 13) (rank:13, 12, 12)
(b) Distractors of different categories

(rank:4, 4, 1) (rank:10, 11, 11) (rank:5, 1, 4)
(d) Distractors with "little boy"

(rank:11, 9, 8) (rank:7, 10, 10) (rank:9, 6, 2)
(c) Distractors with "boy"

(rank:2, 2, 5) (rank:3, 3, 7) (rank:6, 5, 3)
(e) Distractors with "napkin", "basket" and "cup"

Figure 9: Examples of the qualitative results of the proposed MattNet-Mine, MattNet [6] and CM-Att-Erase [3]. For simplicity, we only show the objects with the highest matching score in each image for each method. The results of the proposed MattNet-Mine, MattNet [6] and CM-Att-Erase [3] are marked by yellow, blue and red colours, respectively. The target region in the target image Fig. (a) is bounded by green colour. For each method, we sort the most similar objects of each image by their similarity. The ranking of the objects are shown below the boxes (smaller rank means larger similarity). Best viewed on screen.

Expression: *The large pillow that is under the lying dog that is in the bed.*

(rank:3, 3, 2)
(a) The image with the target "pillow".

(rank:13, 12, 12) (rank:10, 10, 13) (rank:8, 9, 10)
(b) Distractors of different categories

(rank:12, 11, 9) (rank:2, 6, 8) (rank:11, 13, 11)
(c) Distractors with "pillow"

(rank:6, 5, 7) (rank:9, 7, 5) (rank:5, 4, 4)
(d) Distractors with "large pillow"

(rank:1, 1, 3) (rank:4, 2, 1) (rank:7, 8, 6)
(e) Distractors with "pillow", "dog" and "bed"

Figure 10: A failure case of the proposed MattNet-Mine, MattNet [6] and CM-Att-Erase [3]. For simplicity, we only show the objects with the highest matching score in each image for each method. The results of the proposed MattNet-Mine, MattNet [6] and CM-Att-Erase [3] are marked by yellow, blue and red colours, respectively. The target region in the target image Fig. (a) is bounded by green colour. For each method, we sort the most similar objects of each image by their similarity. The ranking of the objects are shown below the boxes (smaller rank means larger similarity). Best viewed on screen.