1 Detailed Analysis on Phrase Localization

1.1 With or Without Regression

Figure 1 compares the results with and without bounding box regression. We use the model of `300-16(-4096)` to generate the regressor (explained in our paper in Sec.5.2). Figure 1a shows the successful cases. The regression is effective especially for the frequently appeared categories in training data such as `person` and `dog` because the accurate regressor can be learned by using many examples. The regression was succeeded for several uncommon categories such as `potter` and `gondola`; the reason is that regressor can be shared with other common categories, e.g., `person` and `boat` regressor can be used for `potter` and `gondola`, respectively. Figure 1b shows the failure cases, which include the categories with ambiguous boundary (e.g., `sidewalk` and `mud`). The regressor does not work for such categories. In addition, if the category is not frequently appeared in training data, the regressor moves the bounding box into the wrong direction. Future work includes automatically determining whether to perform bounding box regression or not.

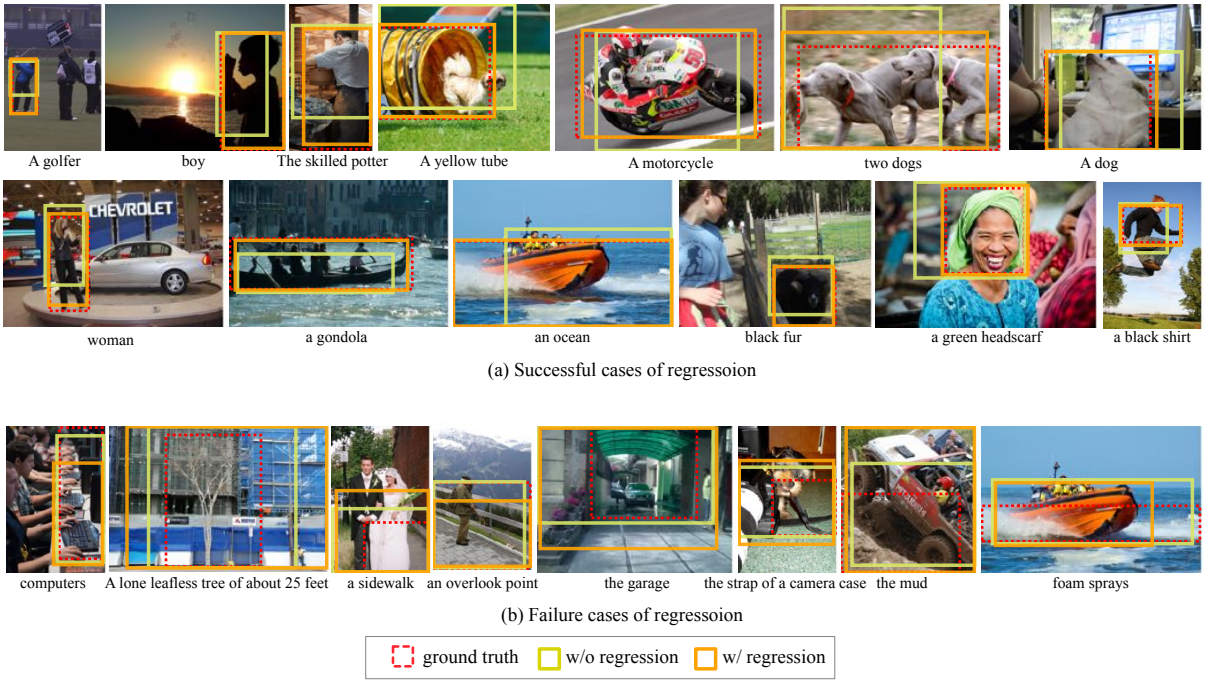

Figure 1: Phrase localization results with and without the bounding box regression. We visualize the ground truth bounding box, the result without the regression, and the result with NPA in red, yellow, and orange, respectively.

1.2 With or Without Negative Phrase Augmentation

Table 1 shows the phrase localization performance with and without negative phrase augmentation (NPA). It shows that the phrase localization performance is not improved by training with NPA. As explained in our paper, it is due to the difference between the phrase localization and object detection tasks; phrase localization assumes there is only one relevant object in the image while object detection places no assumption on the number of objects. Because of this, in the phrase localization task, we can benefit from NPA only when confusing objects appear in the single image. Figure 2a shows such cases: e.g., when two persons appear in the same image, the method with NPA can select the appropriate person that is relevant to the query. However, since such cases are rare in the Flickr30k Entities dataset, NPA does not contribute to performance. Figure 2b shows the failure cases of NPA. The method with NPA tends to predict the small bounding box in which other objects do not appear. The reason is that NPA cannot handle highly overlapped objects appropriately. For example, in the third example in Fig. 2, the `sand` region may have high scores for the `deer`. If there are many such cases in the validation set, the `sand` is added to hard negative phrase for the `deer`. The `sand` classifier thus predicts low score to the regions that are overlapped with the `deer`. Therefore, the method with NPA tends to predict the small box that contains only the (part of) relevant object. This is the limitation of NPA and causes the accuracy decrease in phrase localization task.

Method	People	Clothing	Body	Animals	Vehicles	Inst.	Scene	Other	All
w/o NPA	**78.17**	**61.99**	**35.25**	74.41	**76.16**	56.69	68.07	**47.42**	**65.21**
w/ NPA	77.13	60.06	33.86	**76.76**	73.55	**58.60**	**68.94**	45.28	64.09

Table 1: Comparison of Flickr30k Entities phrase localization performance with and without NPA.

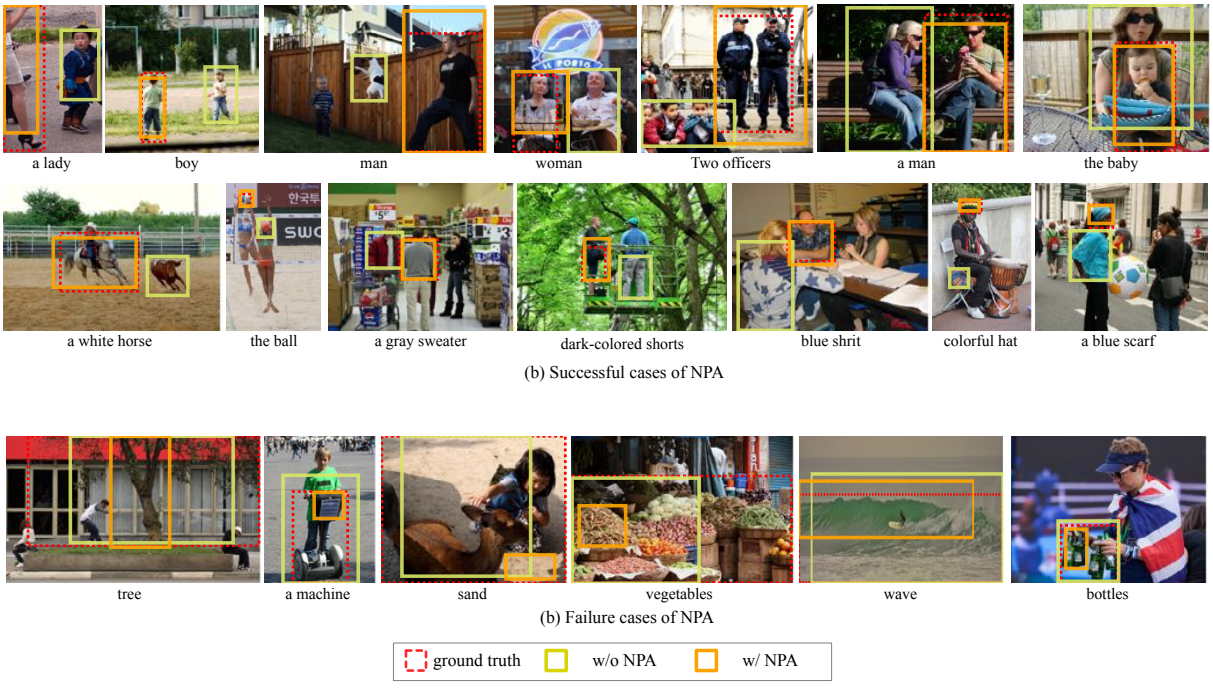

Figure 2: Phrase localization results with and without bounding box regression. We visualize ground truth bounding box, the result without NPA, and the result with NPA in red, yellow, and orange, respectively.

1.3 Ablation Studies

We here present the detailed analysis of our approach on the Flickr30k Entities phrase localization task and quantify our architectural design decisions. For the simplicity, in the comparison of the region proposal and text embedding, we used the pretrained Faster R-CNN model trained on the COCO object detection and finetuned it for phrase localization task. The bounding box regression and NPA are not used in this experiments.

Pretraining. Table 2 compares three pretrained models trained on 1) ImageNet classification, 2) PASCAL, and 3) COCO object detection. In addition, we pretrain the whole model including detector generator using Visual Genome dataset after initial pretraining of 1)–3). The results show that there is more than 4% difference in accuracy between simply using ImageNet pretrained model and pretraining on COCO and Visual Genome. Since Flickr30k Entities dataset does not contain many training examples for each object category, pretraining Faster R-CNN with large object detection datasets is important. Training on the Visual Genome dataset further improves the performance because it contains a much larger number of categories than COCO dataset and detector generator is also pretrained on such rich data.

Pretrained model	VG pretrain?	People	Clothing	Body	Animals	Vehicles	Inst.	Scene	Other	All
ImageNet		74.98	57.34	28.12	71.88	70.93	50.32	67.45	40.34	60.97
ImageNet	✓	76.30	58.30	27.72	74.61	69.19	56.06	**69.07**	43.93	62.76
PASCAL		75.87	58.00	30.69	74.80	73.26	**59.87**	66.52	42.58	62.19
PASCAL	✓	77.06	61.51	34.06	77.15	69.48	57.96	67.95	46.72	64.44
COCO		77.26	60.19	33.86	**75.78**	75.29	56.69	66.83	46.01	64.08
COCO	✓	**78.17**	**61.99**	**35.25**	74.41	**76.16**	56.69	68.07	**47.42**	**65.21**

Table 2: Comparison of different **pre-training** strategies on Flickr30k Entities phrase localization.

Region proposal. Table 3 compares three region proposal approaches: 1) selective search (Uijlings et al., 2013), 2) region proposal network (RPN) trained on COCO dataset, which is frozen during the training of phrase localization, and 3) RPN finetuned on phrase localization task. The number of regions is 2000 for the selective search following (Girshick et al., 2014) and 300 for the RPN following (Ren et al., 2015). In addition, we compared two region sampling strategies: random sampling used in (Girshick, 2015; Ren et al., 2015) and online hard example mining (OHEM) (Shrivastava et al., 2016). The results show that the RPN finetuned for phrase localization task generates much higher quality region proposals than others (12.41% increase in accuracy compared to the selective search), which demonstrates that learning region proposals play an important role in the phrase localization. OHEM further improved the accuracy by 1.56%.

Region proposal	OHEM?	People	Clothing	Body	Animals	Vehicles	Inst.	Scene	Other	All
Selective search		60.65	44.55	23.96	65.04	68.90	36.94	55.71	34.43	50.11
RPN (COCO pretrained)		71.29	44.82	17.23	70.90	67.44	42.04	63.23	38.26	55.94
RPN (COCO pretrained)	✓	72.12	42.62	16.24	71.88	67.15	44.59	65.09	36.45	55.71
RPN (Flickr30k finetuned)		75.90	58.74	28.32	72.66	73.55	55.41	65.34	44.88	62.52
RPN (Flickr30k finetuned)	✓	**77.26**	**60.19**	**33.86**	**75.78**	**75.29**	**56.69**	**66.83**	**46.01**	**64.08**

Table 3: Comparison of different **region proposals** and region sampling strategies on Flickr30k Entities phrase localization.

Text embedding. Table 4 compares five text embedding vectors: 1) Word2Vec (Mikolov et al., 2013) trained on Google News dataset[1], which is used in our paper, 2) Word2Vec trained on Flickr tags[2] (Li et al., 2015), 3) Hybrid Gaussian-Laplacian mixture model (HGLMM) (Klein et al., 2015), which is used in (Plummer et al., 2015, 2017; Wang et al., 2016), 4) Skip-thought vector (combine-skip

[1] https://code.google.com/archive/p/word2vec/

[2] the model is provided by the author of (Dong et al.)

model)[3] (Kiros et al., 2015), and 5) Long-short term memory (LSTM) that encodes a phrase into a vector in the manner described in (Chen et al., 2017; Rohrbach et al., 2016), which is learned jointly with other components of Query-Adaptive R-CNN. The second column of Table 4 shows the dimension of the text embedding vector. This result shows that the performance is not much affected by the choice of the text embedding. The mean pooling of Word2Vec performs the best despite its simplicity.

Text embedding	dim	People	Clothing	Body	Animals	Vehicles	Inst.	Scene	Other	All
Word2Vec avg.	300	**77.26**	60.19	33.86	75.78	75.29	56.69	66.83	**46.01**	**64.08**
Word2Vec avg. (Flickr tags)	300	75.36	60.19	31.88	75.00	**78.78**	55.41	**68.39**	44.64	63.19
HGLMM	15000	**77.26**	**61.34**	32.28	75.00	68.31	**63.06**	67.33	45.25	63.96
Skip-thought	4800	77.06	59.89	**34.65**	**79.88**	73.55	57.32	68.01	45.28	64.06
LSTM	1000	75.45	58.96	28.71	74.61	75.58	56.05	66.71	29.23	62.36

Table 4: Comparison of different **text embedding** on Flickr30k Entities phrase localization.

[3]We use the implementation and pre-trained model provided in https://github.com/ryankiros/skip-thoughts

2 Additional Examples of Negative Phrase Augmentation

Figure 3, 4, and 5 show additional examples of the negative phrase augmentation (corresponds to Fig. 4 in our paper). There are many false alarms between the confusing categories such as the animal (`zebra`, `bear`, and `giraffe`), person (`skier` and `child`), and vehicle (`boat`, `train`, and `bus`) without NPA, which are successfully discarded by training with NPA.

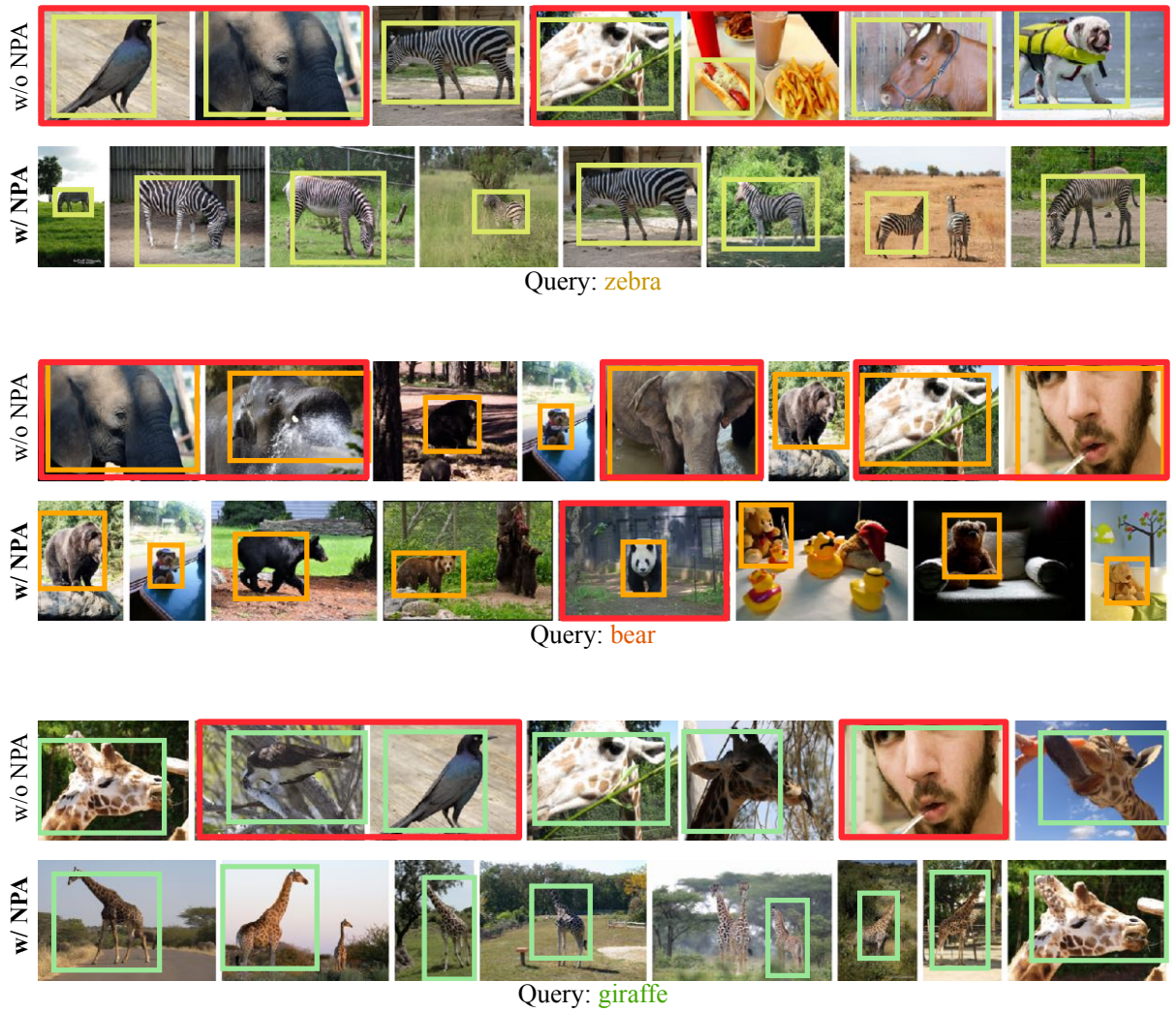

Figure 3: Qualitative results with and without NPA. Top-ranked retrieved results are shown and false alarms are depicted with red border.

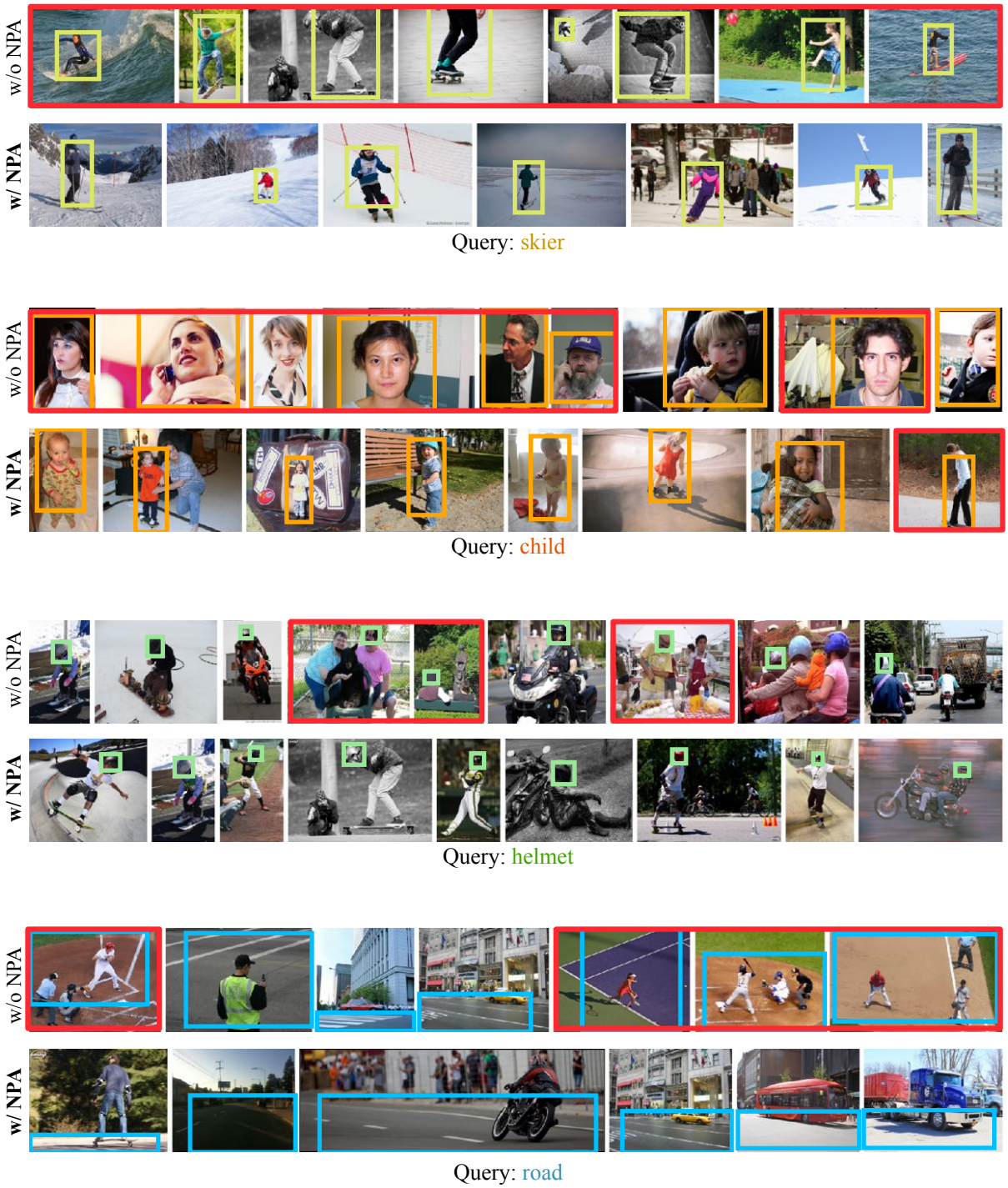

Figure 4: Qualitative results with and without NPA. Top-ranked retrieved results are shown and false alarms are depicted with red border.

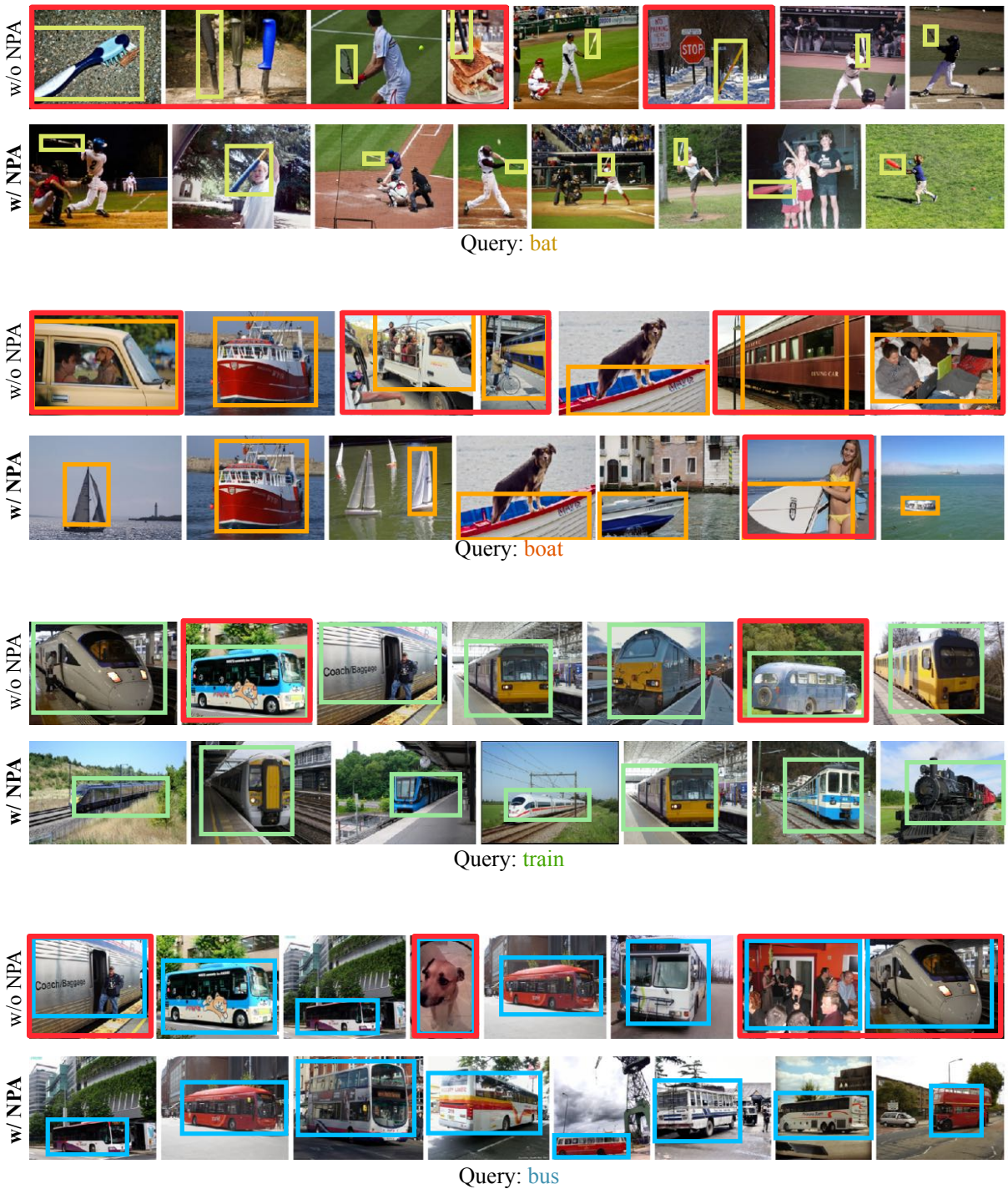

Figure 5: Qualitative results with and without NPA. Top-ranked retrieved results are shown and false alarms are depicted with red border.

3 Additional Examples of Open-Vocabulary Object Retrieval and Localization

Figure 6 shows the additional examples of object retrieval and localization (corresponds to Fig. 6 in our paper). Instead of the ILSVRC dataset used in our paper, we here used the Microsoft COCO dataset (Lin et al., 2014) (40504 images from the validation set) that contains a wider variety of concepts. These results demonstrate that our system can accurately search the wide variety of objects specified by the natural language query.

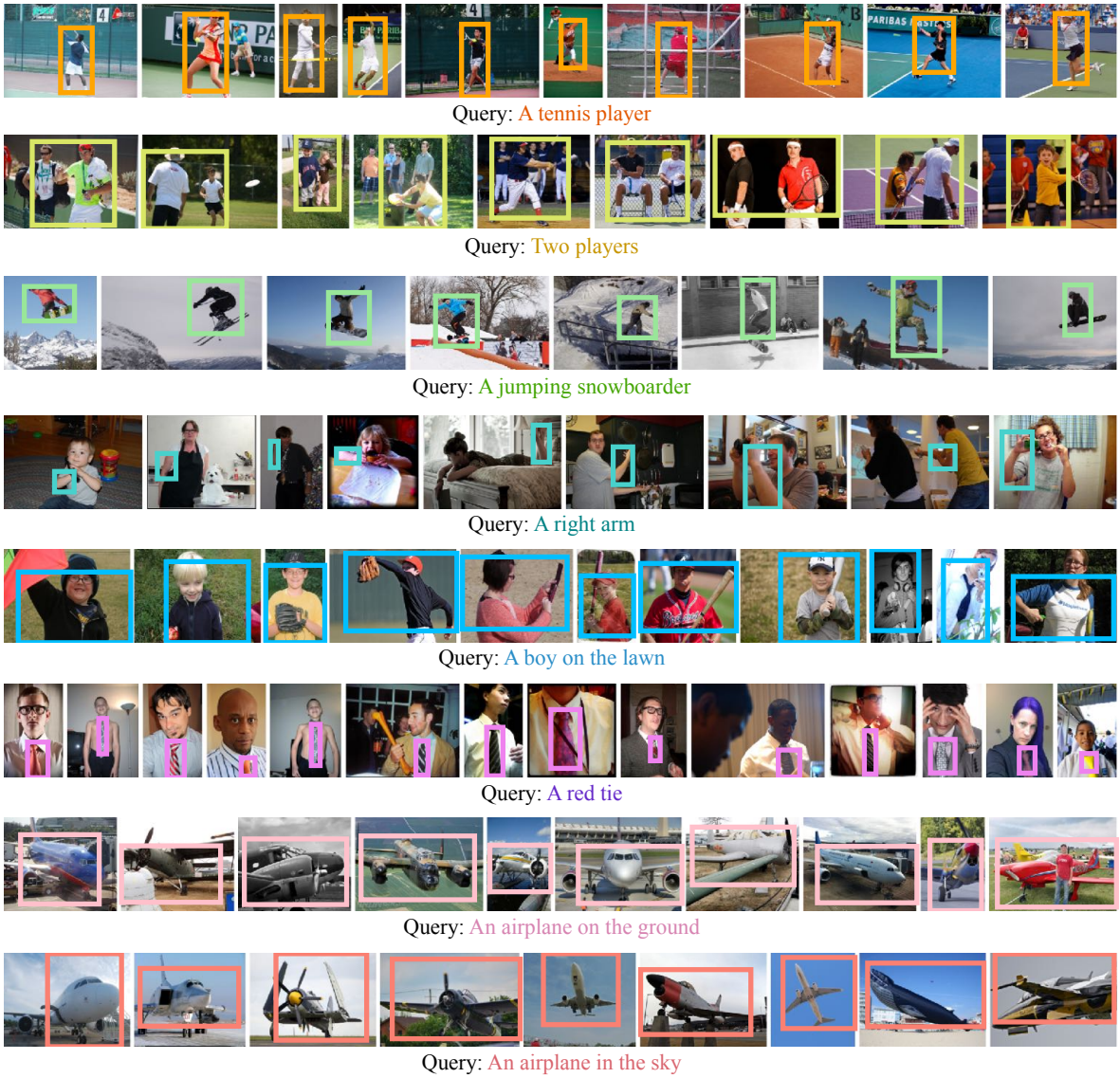

Figure 6: Retrievals from COCO validation set. Top-ranked retrieval results for each query are shown.

References

Kan Chen, Rama Kovvuri, and Ram Nevatia. 2017. Query-guided Regression Network with Context Policy for Phrase Grounding. In *ICCV*.

Jianfeng Dong, Xirong Li, and Cees G. M. Snoek. Word2visualvec: cross-media retrieval by visual feature prediction. *CoRR*, abs/1604.0.

Ross Girshick. 2015. Fast r-cnn. In *ICCV*.

Ross Girshick, Jeff Donahue, Trevor Darrell, U C Berkeley, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *NIPS*.

Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2015. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. In *CVPR*.

Xirong Li, Shuai Liao, Weiyu Lan, Xiaoyong Du, and Gang Yang. 2015. Zero-shot Image Tagging by Hierarchical Semantic Embedding. In *SIGIR*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft coco: common objects in context. In *ECCV*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

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 *ICCV*.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *International Journal of Computer Vision*, 123(1):74–93.

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

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

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. 2016. Training region-based object detectors with online hard example mining. In *CVPR*.

Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. 2013. Selective Search for Object Recognition. *IJCV*, 104(2):154–171.

Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *CVPR*.