Supplementary Material for "Box-based Refinement for Weakly Supervised and Unsupervised Localization Tasks"

Eyal Gomel
Tel Aviv University
eyalgomel12@gmail.com

Tal Shaharbany
Tel Aviv University
shaharabany@mail.tau.ac.il

Lior Wolf
Tel Aviv University
wolf@cs.tau.ac.il

Abstract

This appendix presents visual results that demonstrate the effectiveness of our refined models g^h and f^h in various tasks, including weakly supervised and unsupervised localization, What-is-where-by-looking, and unsupervised single object discovery. By building upon existing models g and f, we have showcased improvements in output localization maps and bounding boxes.

Our comprehensive comparisons span multiple datasets, including MS-COCO14 [7], Visual-Genome [6], Flickr30K [8], ReferIt [2, 5], PASCAL-VOC07 [3], PASCAL-VOC12 [4], and MS-COCO20K [7]. These comparisons serve to highlight the adaptability and robustness of our refined models across different tasks and datasets. The visual results provide strong evidence of our models' superiority in generating more accurate localization maps and bounding boxes compared to their base models.

The code and scripts for reproducing the paper's results are attached to this supplementary.

A. Weakly supervised phrase-grounding visual results

We present visual outcomes of our model, g^h, which is built upon the previously published model g by [9]. We compare the localization maps and bounding box outputs generated by both models and evaluate each bounding box against the ground truth. We showcase the results for models trained on the MS-COCO14 [7] and Visual-Genome [6] datasets. For each model, we display visualizations on the Flickr30K[8], ReferIt [2, 5], and Visual-Genome [6] datasets. Figures 1, 2, 3 illustrate the results for the MS-COCO-based model, while the outcomes for the VG-based model can be found in Figures 4, 5, 6.

B. What is where by looking visual results

We present visual outcomes for the What-is-where-by-looking task using our improved model g^h, which is derived from the previously published model g by [9]. We compare the localization maps generated by both models, using the same image but different phrases. In Figure 7, we display the results for the Flickr30K[8] dataset, with models g and g^h trained on the MS-COCO14 [7] dataset.

C. Unsupervised single object discovery visual results

In the context of the unsupervised single object discovery task, we display visualizations of our model f^h, which is based on the DINO[1] model f. We compare our findings with those of LOST[10] and TokenCut[11]. For each comparison, we showcase the output attention map and the output bounding box. Additionally, we display CAD-based bounding boxes, derived from both our refined model f^h and the original model f, if available. For each method, we exhibit results on the PASCAL-VOC07 [3], PASCAL-VOC12 [4], and MS-COCO20K[7] datasets. The outcomes for the LOST model can be found in Figures 8,9,10, while the TokenCut model results are illustrated in Figures 11, 12, 13.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 1

[2] Kan Chen, Rama Kovvuri, and Ram Nevatia. Query-guided regression network with context policy for phrase grounding. In *ICCV*, 2017. 1, 4, 7

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 Results. pascal-network.org/challenges/VOC/voc2007. 1, 10, 13

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012. 1, 11, 13

[5] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. The IAPR TC-12 benchmark: A new evaluation resource for visual information systems. In *International workshop ontoImage*, volume 2, 2006. 1, 4, 7

[6] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 1, 5, 6, 7, 8

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, volume 8693 of *LNCS*, pages 740–755, 2014. 1, 3, 4, 5, 9, 10, 12

[8] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 1, 3, 6, 9

[9] Tal Shaharabany, Yoad Tewel, and Lior Wolf. What is where by looking: Weakly-supervised open-world phrase-grounding without text inputs. In *NeurIPS*, 2022. 1, 3, 4, 5, 6, 7, 8, 9

[10] Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. In *BMVC*, 2021. 1, 10, 11

[11] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L. Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *CVPR*, 2022. 1, 12, 13

Figure 1. Phrase-grounding results on Flickr30K[8] dataset. Model g^h was trained on MS-COCO14[7] dataset. (a) the phrase (b) the input image (c) results (black) for network g [9] compared to ground-truth box (green) (d) same for refined network g^h. (e) same as a (f) same as b (g) same as c (h) same as d

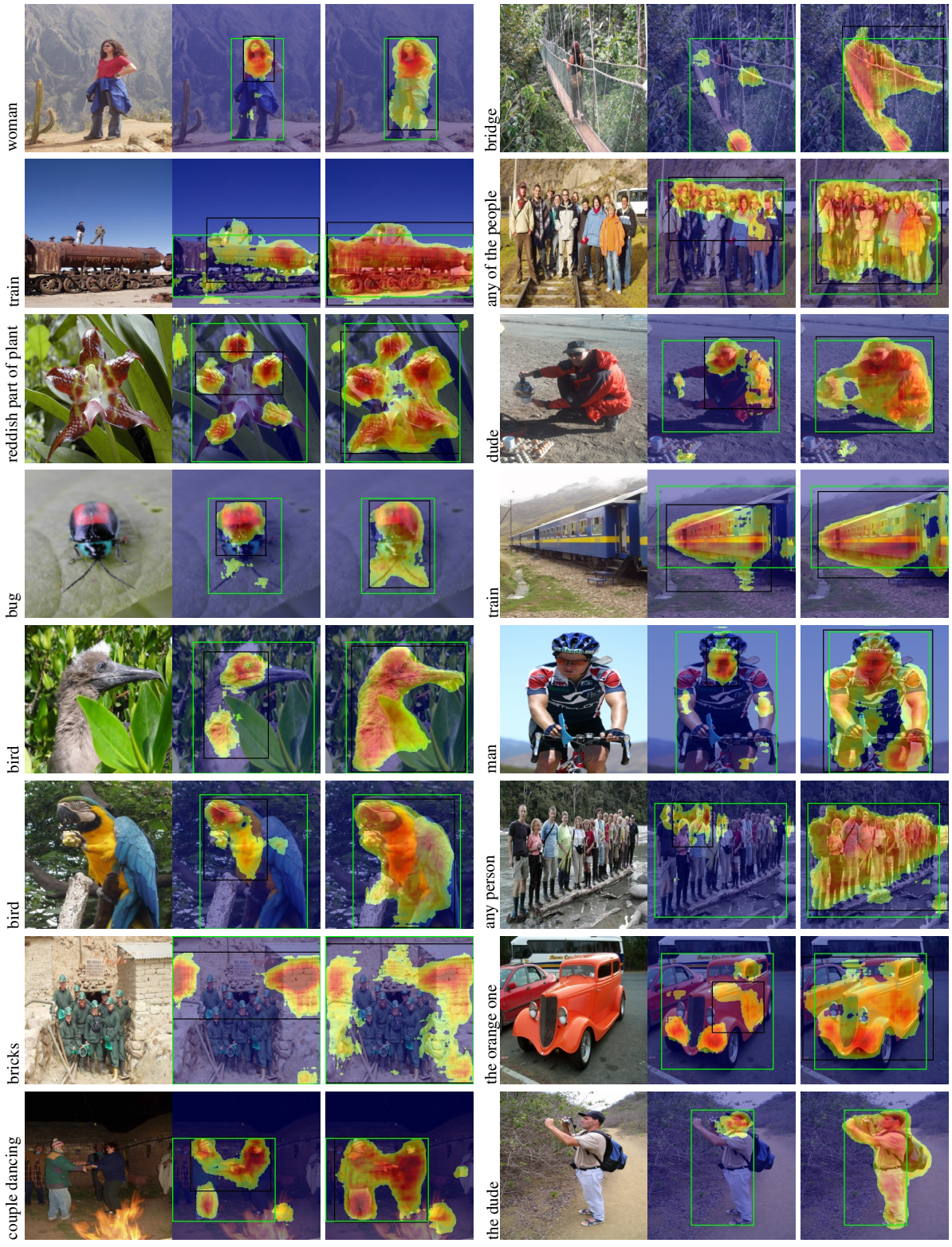


Figure 2. Phrase-grounding results on ReferIt[2, 5] dataset. Model g^h was trained on MS-COCO14[7] dataset. (a) the phrase (b) the input image (c) results (black) for network g [9] compared to ground-truth box (green) (d) same for refined network g^h. (e) same as a (f) same as b (g) same as c (h) same as d

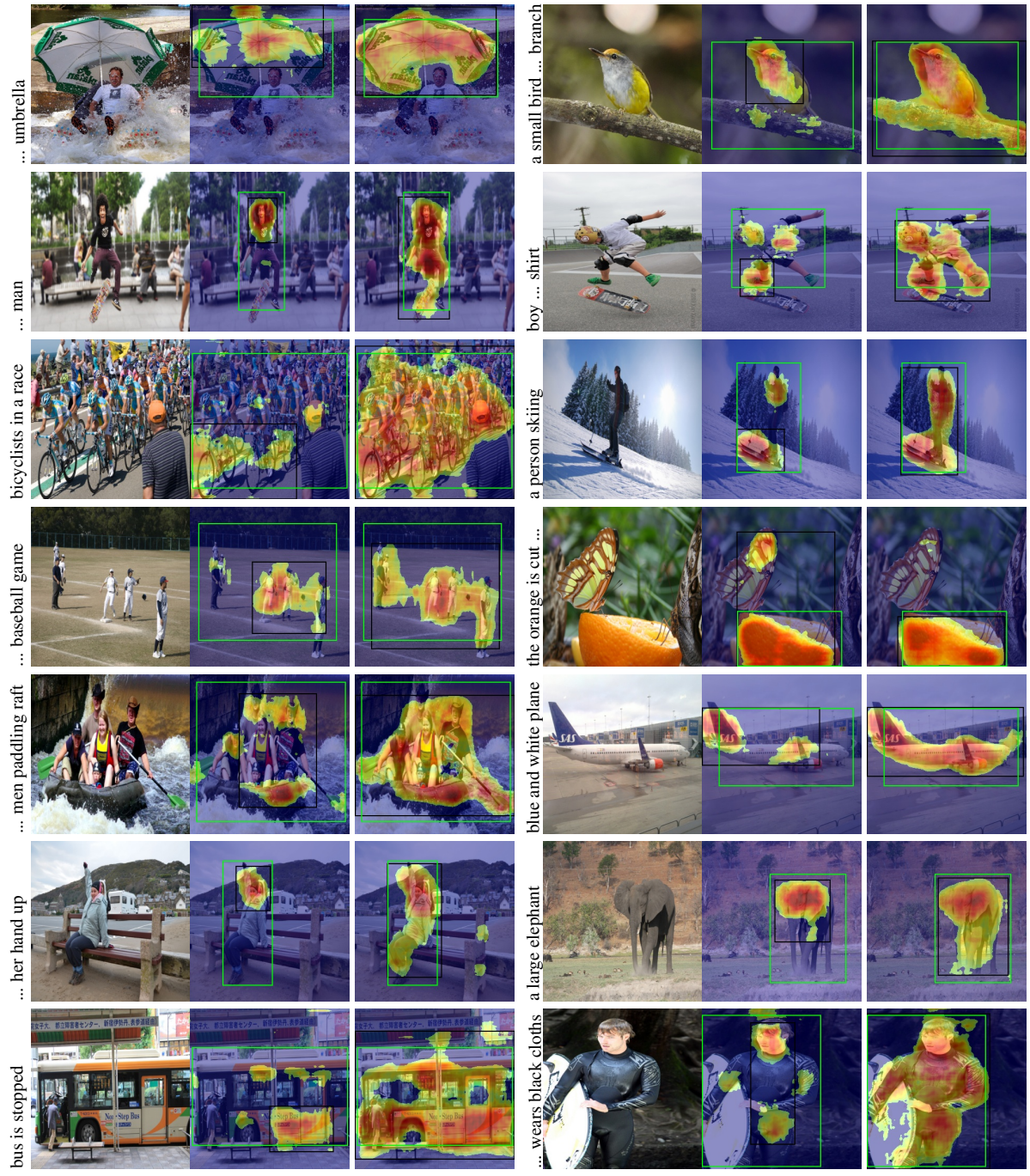


Figure 3. Phrase-grounding results on Visual Genome [6] dataset. Model g^h was trained on MS-COCO14[7] dataset. (a) the phrase (b) the input image (c) results (black) for network g [9] compared to ground-truth box (green) (d) same for refined network g^h. (e) same as a (f) same as b (g) same as c (h) same as d

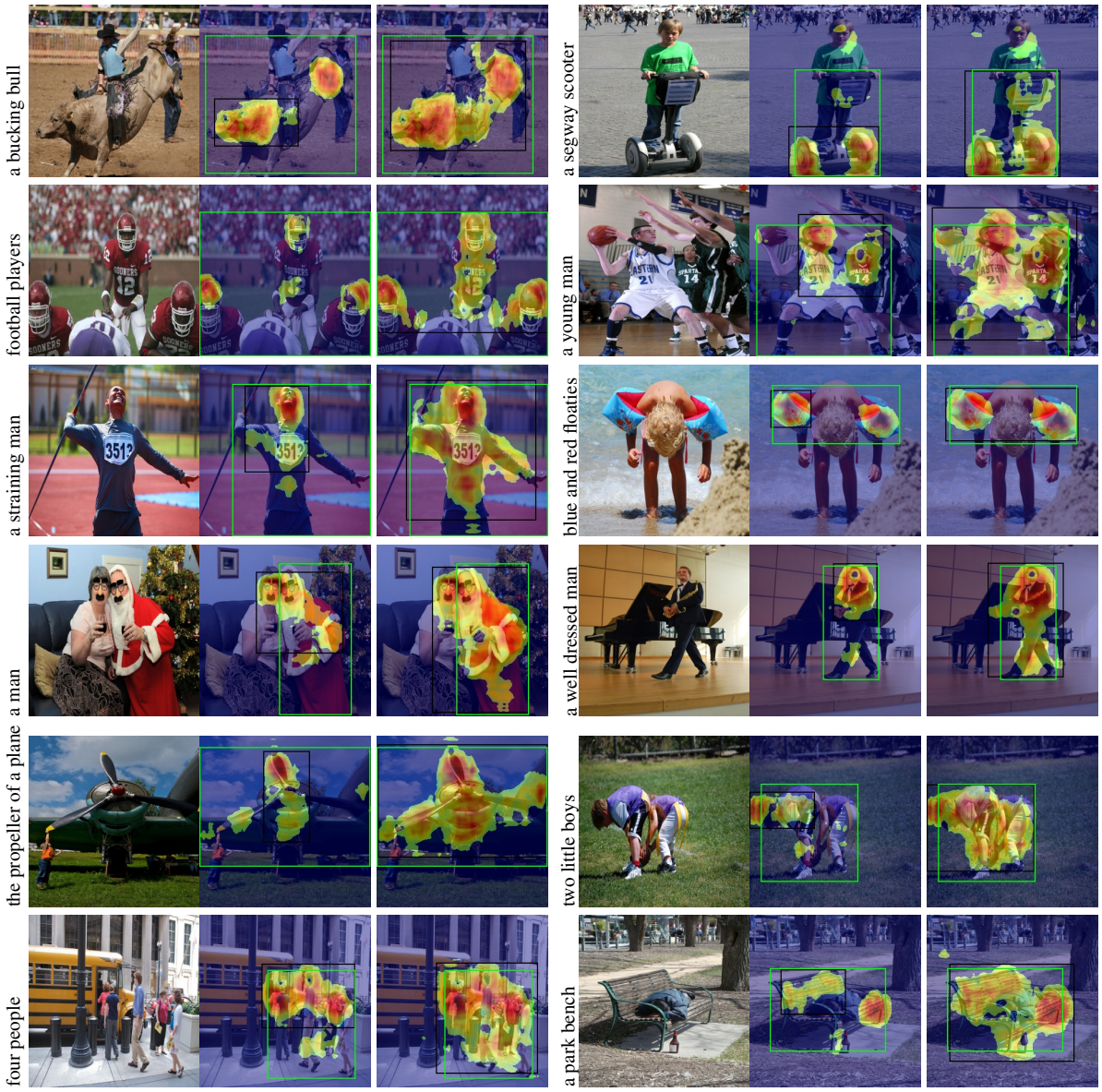


Figure 4. Phrase-grounding results on Flickr30K[8] dataset. Model g^h was trained on Visual Genome [6] dataset. (a) the phrase (b) the input image (c) results (black) for network g [9] compared to ground-truth box (green) (d) same for refined network g^h. (e) same as a (f) same as b (g) same as c (h) same as d

Figure 5. Phrase-grounding results on ReferIt[2, 5] dataset. Model g^h was trained on Visual Genome [6] dataset. (a) the phrase (b) the input image (c) results (black) for network g [9] compared to ground-truth box (green) (d) same for refined network g^h. (e) same as a (f) same as b (g) same as c (h) same as d

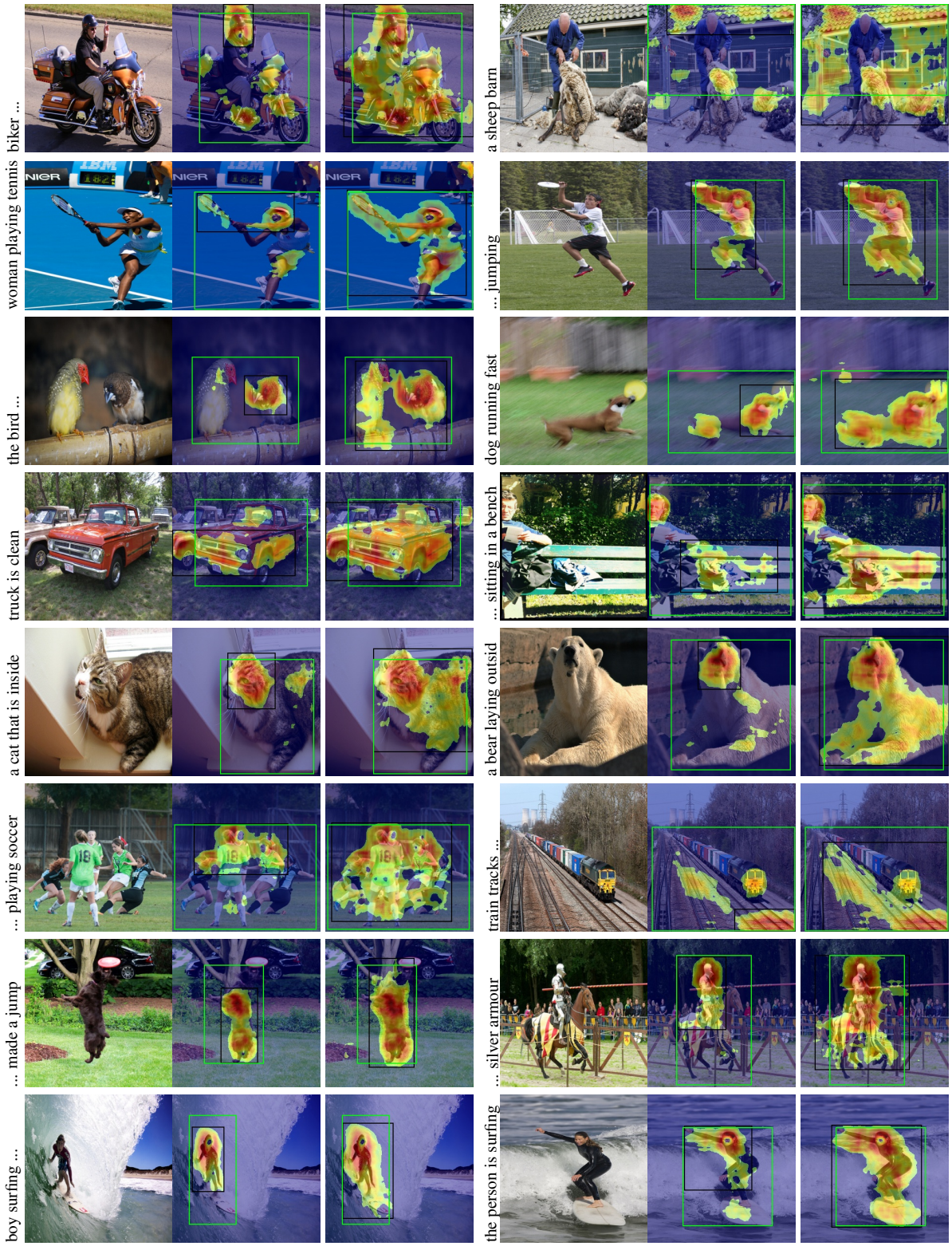
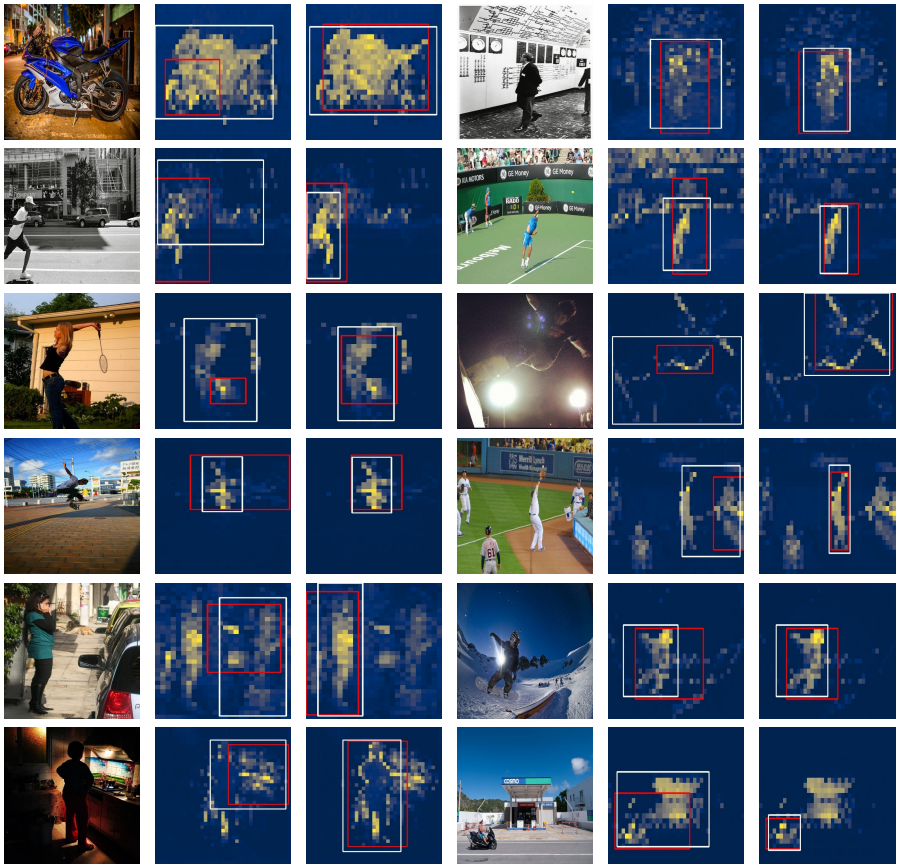


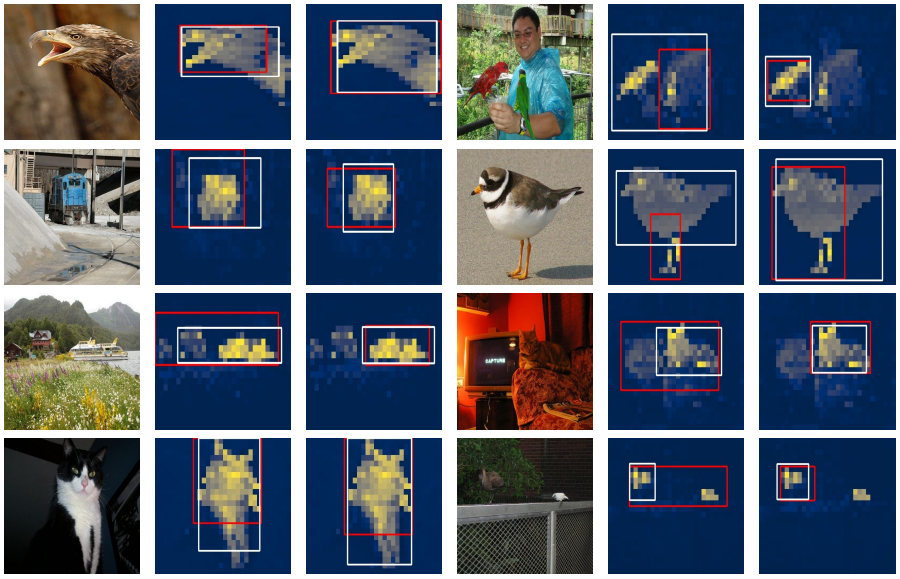
Figure 6. Phrase-grounding results on Visual Genome [6] dataset. Model g^h was trained on the same dataset. (a) the phrase (b) the input image (c) results (black) for network g [9] compared to ground-truth box (green) (d) same for refined network g^h. (e) same as a (f) same as b (g) same as c (h) same as d

Figure 7. What-is-where-by-looking results on Flickr30K[8] dataset. Model g^h was trained on MS-COCO14[7] dataset. (a) the input image (b) results for network g [9] (c) results for network g^h (d-e) same as b-c, using different phrase



(a) (b) (c) (d) (e) (f)

Figure 8. Single object discovery results on MS-COCO14[7] dataset. (a) the input image (b) the inverse degree of the LOST [10]; the red bounding box is directly from LOST, the white is the prediction of CAD trained on top of it (c) same with our refined model f^h and LOST (d) same as a (e) same as b (f) same as c

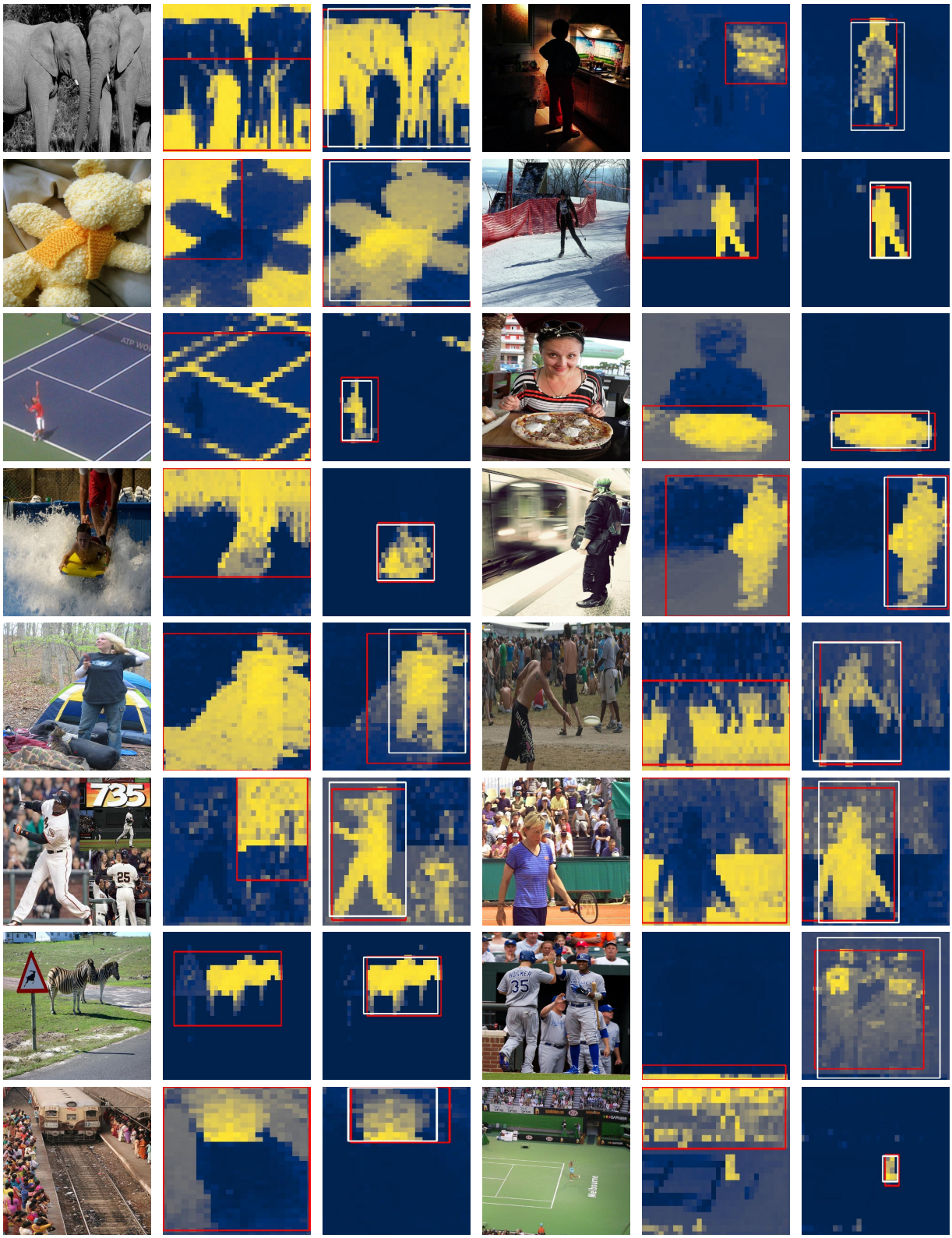


(a) (b) (c) (d) (e) (f)

Figure 9. Single object discovery results on PASCAL-VOC07[3] dataset. (a) the input image (b) the inverse degree of the LOST [10]; the red bounding box is directly from LOST, the white is the prediction of CAD trained on top of it (c) same with our refined model f^h and LOST (d) same as a (e) same as b (f) same as c

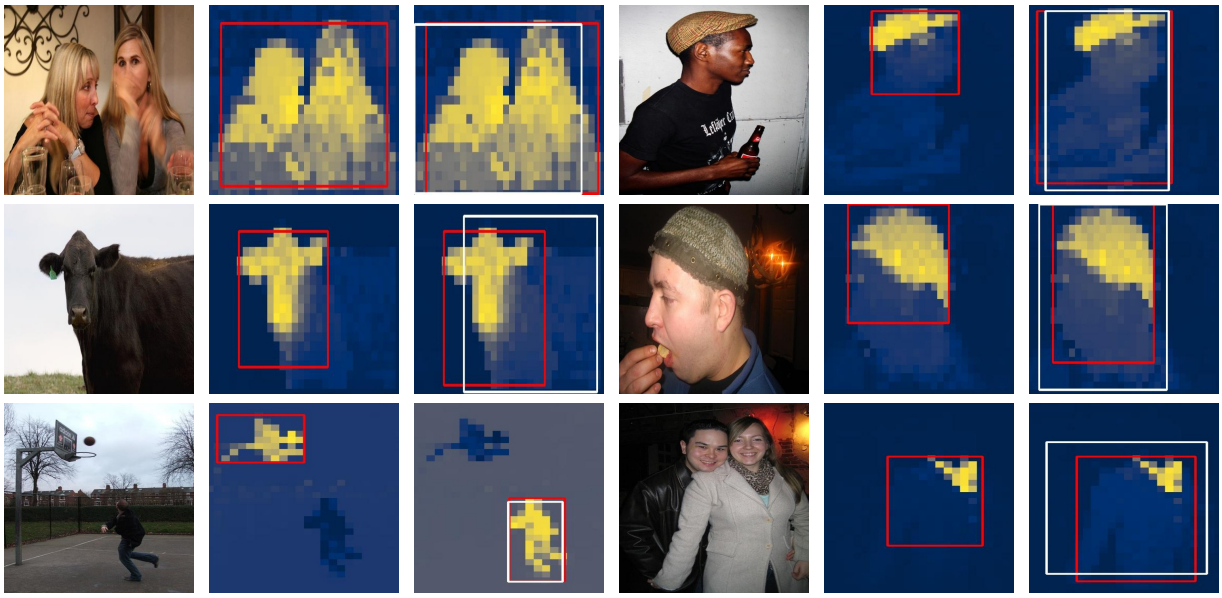
(a) (b) (c) (d) (e) (f)

Figure 10. Single object discovery results on PASCAL-VOC12[4] dataset. (a) the input image (b) the inverse degree of the LOST [10]; the red bounding box is directly from LOST, the white is the prediction of CAD trained on top of it (c) same with our refined model f^h and LOST (d) same as a (e) same as b (f) same as c



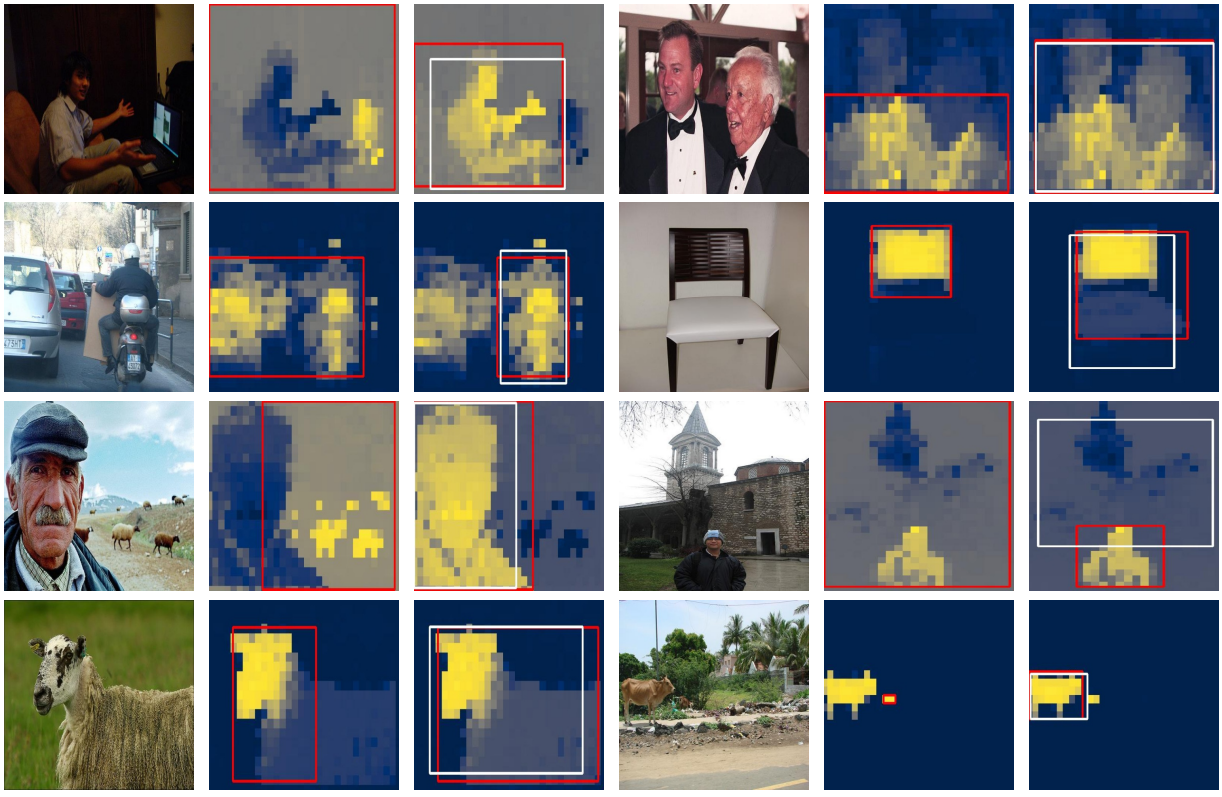
(a) (b) (c) (d) (e) (f)

Figure 11. Single object discovery results on MS-COCO14[7] dataset. (a) the input image (b) the eigenvector attention of the TokenCut [11]; the red bounding box is directly from TokenCut (the CAD model was not released and is not shown) (c) same with our refined model f^h and TokenCut, the white bounding box is the prediction of CAD trained on top of f^h (d) same as a (e) same as b (f) same as c



(a) (b) (c) (d) (e) (f)

Figure 12. Single object discovery results on PASCAL-VOC07[3] dataset. (a) the input image (b) the eigenvector attention of the TokenCut [11]; the red bounding box is directly from TokenCut (the CAD model was not released and is not shown) (c) same with our refined model f^h and TokenCut, the white bounding box is the prediction of CAD trained on top of f^h (d) same as a (e) same as b (f) same as c



(a) (b) (c) (d) (e) (f)

Figure 13. Single object discovery results on PASCAL-VOC12[4] dataset. (a) the input image (b) the eigenvector attention of the TokenCut [11]; the red bounding box is directly from TokenCut (the CAD model was not released and is not shown) (c) same with our refined model f^h and TokenCut, the white bounding box is the prediction of CAD trained on top of f^h (d) same as a (e) same as b (f) same as c