# Supplementary to Localized Vision-Language Matching for Open-vocabulary Object Detection

María A. Bravo, Sudhanshu Mittal, and Thomas Brox

Department of Computer Science
University of Freiburg, Germany
{bravoma, mittal, brox}@cs.uni-freiburg.de

## Appendix

## A VAW dataset

**Visual Attributes in the Wild (VAW) dataset [2]** We use the training, validation and test set of images as defined with the proposed dataset [2]. The dataset contains 58,565 images for training, 3,317 images for validation, and 10,392 images for testing. We define the splits for known and novel classes taking approximately 20% of the total classes (2260) to be novel, resulting in 1792 known and 468 novel classes. We make sure that all known and novel classes from COCO split are kept in the same subset for VAW splits. After removing images with no known annotations from the training and splitting into known and novel classes, there are 54,632 images for training spanning over 1790 known classes, 818 known / 200 novel classes for the validation set, and 1020 known / 297 novel classes for the test set. This dataset is much more challenging as compared to COCO since it contains fine-grained classes with a long-tailed distribution. It not only contains more classes as compared to the COCO benchmark, but also poses additional challenges like plural versions defined as different classes, *e.g.*kites vs kite. In the LSM phase, we use the captions from **Visual Genome Region Descriptions [1]** which contain 108,077 images with a total of 4,297,502 region descriptions. We combine these region descriptions for every image to have a single caption per image.

**VAW dataset results.** LocOv successfully generalizes to the VAW benchmark. Table 1 shows the comparison of our approach to both STT-ZSD and

| Method | Novel (297) | | | Known (1020) | | | Generalized (2060) | | |
|---|---|---|---|---|---|---|---|---|---|
| | AP | $AP_{50}$ | $AP_{75}$ | AP | $AP_{50}$ | $AP_{75}$ | AP | $AP_{50}$ | $AP_{75}$ |
| STT-ZSD (Ours) | 0.14 | 0.28 | 0.15 | **1.33** | **2.56** | **1.16** | **0.95** | **1.84** | **0.82** |
| OVR [3] | 0.59 | 1.27 | 0.45 | 0.92 | 2.08 | 0.72 | 0.70 | 1.57 | 0.54 |
| LocOv (Ours) | **0.67** | **1.42** | **0.59** | 1.21 | 2.31 | 1.11 | 0.91 | 1.77 | 0.81 |

Table 1: Comparing open-vocabulary object detection results on the VAW test set.

OVR baselines on the test set. Our method improves consistently over the other two methods for the novel classes, showing that it can scale to more challenging settings with long-tailed distribution and large number of classes.

## B Ablation Experiments

Table 2: Comparison of the different stages of the model on the novel object detection. The table also shows different configurations of model update in the STT stage by freezing parts of the backbone network

| LSM | STT | Freezing blocks | | | Novel (17) | | | Known (48) | | | Generalized | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1-4 | 1-3 | 1-2 | AP | $AP_{50}$ | $AP_{75}$ | AP | $AP_{50}$ | $AP_{75}$ | AP | $AP_{50}$ | $AP_{75}$ |
| ✓ | ✓ | | | ✓ | **17.17** | 30.86 | **16.78** | 30.79 | 50.68 | 32.21 | **26.14** | **43.80** | **27.05** |
| ✓ | ✓ | | ✓ | | 16.77 | **30.91** | 16.24 | 30.10 | 49.71 | 31.14 | 25.44 | 43.14 | 25.98 |
| ✓ | ✓ | ✓ | | | 15.96 | 29.09 | 15.59 | 29.14 | 48.50 | 30.63 | 24.82 | 41.99 | 25.73 |
| ✓ | | | | ✓ | 0.73 | 1.89 | 0.37 | 0.82 | 2.06 | 0.48 | 0.89 | 2.27 | 0.52 |
| | ✓ | | | ✓ | 0.21 | 0.31 | 0.21 | **33.23** | **53.43** | **35.03** | 24.38 | 39.19 | 25.72 |

*Two-stage model performs best.* In Table 2, we show the extended performance of our method using single stage, either LSM or STT, and fine-tuning different sets of the backbone weights during the STT stage. The last two rows of Table 2 consider our method using only the STT stage (same as our baseline STT-ZSD from Section 4.2 in the main paper) and using only the LSM stage. Individual stage models are not able to detect novel objects well, which shows that both stages are fundamental for the detection of novel objects. We further compare the performance of different model configurations by freezing different number of blocks of the backbone network during the STT stage. Our results show that only freezing the first two blocks and the projection layer leads to the best configuration for the STT. In conclusion we can observe two main results: first, using both stages is crucial to detect novel objects. Second, freezing the backbone weights of the 1st and 2nd ResNet blocks during the STT stage results in the best configuration for both, novel and known, performances.

*Localized objects matter.* Table 3 presents the impact of using box- vs grid-region features in the LSM stage. We compare our method using grid-region features $R^I_{grid}$, proposed box-region features $R^I_{box}$, and using box-region features from the known ($k$) or novel ($n$) class annotations $R^I_{ann}$. When training the LSM stage, we only consider a fixed amount of image regions to calculate the losses and drop the rest of the regions. To illustrate that the improvement comes from the combination of grid- and box-regions and not simply from more boxes, we trained with an increased number of image regions (100 and 200) for every case explicitly stated in Table 3. Even though increasing the number of regions results

Table 3: Different image regions for the LSM stage. $R^I_{grid}$- grid-regions, $R^I_{box}$- proposed box-regions and $R^I_{ann}$- ground truth box-regions of (k) known or (n) novel objects use during the LSM stage

| Regions | | | Novel (17) | | | Known (48) | | | Generalized | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $R^I_{grid}$ | $R^I_{box}$ | $R^I_{ann}$ | AP | $AP_{50}$ | $AP_{75}$ | AP | $AP_{50}$ | $AP_{75}$ | AP | $AP_{50}$ | $AP_{75}$ |
| 100 | | k+n | 18.2 | 31.6 | 18.2 | 32.5 | 52.7 | 34.0 | 27.9 | 46.0 | 28.8 |
| | | k+n | 16.3 | 28.4 | 15.9 | 32.9 | 53.1 | 34.9 | 27.6 | 45.3 | 28.8 |
| 100 | | k | 14.2 | 26.8 | 13.4 | 30.0 | 50.2 | 31.3 | 24.8 | 42.4 | 25.5 |
| 100 | 100 | | **17.2** | **30.1** | **17.5** | 33.5 | 53.4 | 35.5 | **28.1** | **45.7** | **29.6** |
| 200 | | | 15.5 | 27.1 | 15.4 | 32.2 | 52.1 | 33.9 | 27.1 | 44.5 | 28.2 |
| 100 | | | 14.9 | 25.8 | 15.0 | 31.7 | 51.8 | 33.3 | 26.6 | 43.9 | 27.7 |
| | 200 | | 13.7 | 25.7 | 12.9 | **34.2** | **53.8** | **36.5** | 27.5 | 43.8 | 29.1 |
| | 100 | | 13.4 | 22.8 | 13.4 | 33.9 | 53.7 | 35.8 | 27.0 | 43.3 | 28.5 |

Table 4: Ablation study showing the performance of using of BERT text embeddings vs BERT Model during the LSM stage on COCO validation set.

| BERT Model | BERT Emb. | Novel (17) | | | Known (48) | | | Generalized | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | AP | $AP_{50}$ | $AP_{75}$ | AP | $AP_{50}$ | $AP_{75}$ | AP | $AP_{50}$ | $AP_{75}$ |
| | ✓ | 17.0 | 29.6 | 17.1 | 33.5 | 53.4 | 35.5 | 28.1 | 45.7 | 29.6 |
| | | ±0.3 | ±0.7 | ±0.7 | ±0.1 | ±0.1 | ±0.0 | ±0.0 | ±0.1 | ±0.1 |
| ✓ | ✓ | 16.6 | 29.3 | 16.4 | 32.6 | 52.3 | 34.5 | 27.4 | 45.0 | 28.6 |
| | | ±0.1 | ±0.5 | ±0.2 | ±1.4 | ±1.8 | ±1.5 | ±1.0 | ±1.2 | ±1.2 |

in a better performance the combination of both types of regions proves to be best, showing a complementary behavior. We also considered two oracle experiments (row 1 and 2) using ground-truth box-region features from both known and novel class annotations instead of proposed box-region features. These two experiments improve performance on novel classes showing that object-centered box regions are crucial and the best performance is achieved when combined with additional grid regions (row 1). The additional grid-regions help in capturing the background objects beyond the annotated classes while box-regions focus on precise foreground objects, which improves the image-caption matching.

*Text embedding selection.* To verify the improvement of using simplified text embeddings over BERT Model, we perform three runs using both configurations during the LSM stage. Table 4 shows the mean results over the runs with their standard deviation. Even though the differences between the two model configurations are small, on average using the simple embedding layer of BERT gives a higher performance.

## C Limitations

Visual features of novel object classes are learned during the Localized Semantic Matching stage using image-caption pairs. We notice that such a weak form of

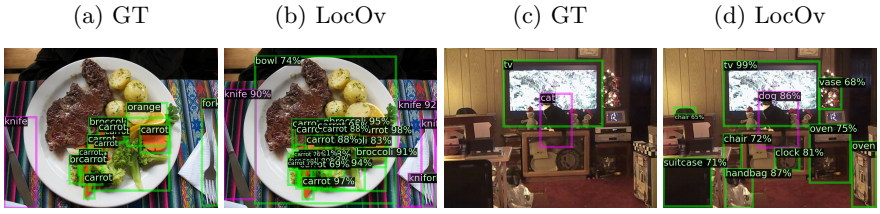


Fig. 1: Failure cases. The method fails to learn fine-grained classification for novel objects. The model confuses between similar classes. For *e.g.* the model sometimes predicts 'fork' as 'knife'(left image) and 'cat' as 'dog'(right image).

supervision is not sufficient to learn fine-grained classification. Similar classes such as 'dog' and 'cat' or 'knife' and 'fork' are often confused, as shown in Figure 1, since they can be used exchangeably in the caption description and they sometimes even co-occur in the image (*e.g.*knife-fork), making the matching process ambiguous. We also observe a clear drop in performance of known object classes when a similar novel object class is detected. A table showing this analysis quantitatively is included in the supplementary.

## D Per Class Performance

Figure 2 presents the difference of AP per class when considering the generalized setup, all classes together, minus the AP for the individual setup, only the novel or only the known classes. Most of the scores present a drop when considering the generalized case. Analyzing cases where this drop is larger than 3.5 AP (the red bars in Figure 2) we can deduce that these classes are mostly confused. Figures 3 and 4 show some qualitative examples of our method. We show the ground truth image with annotations and results using our method for comparison. In Figure 3 we can observe that classes such as bowl and cup are frequently confused, and similar error occurs for classes: fork, knife and spoon. These errors occur due to the fact that these classes look similar or appear together very often. These type of errors are also noticeable between other such classes like cow/sheep/dog and snowboard/skis/skateboard. The class toaster is a special case since it is the class with the least instances present in the dataset (only 9 vs a median of 275), which makes it harder for our method to distinguish this class among the known set and the task becomes harder when considering all 65 classes.

## E Qualitative examples

Figures 3 and 4 show some random qualitative examples of LocOv . Our method is capable of discovering novel classes such as cat, dog, sink, bus with high confidence, specially when there is no ambiguity or similarity with other categories. Similar visual classes such as fork, knife and spoon; cow and sheep; cat and dog;

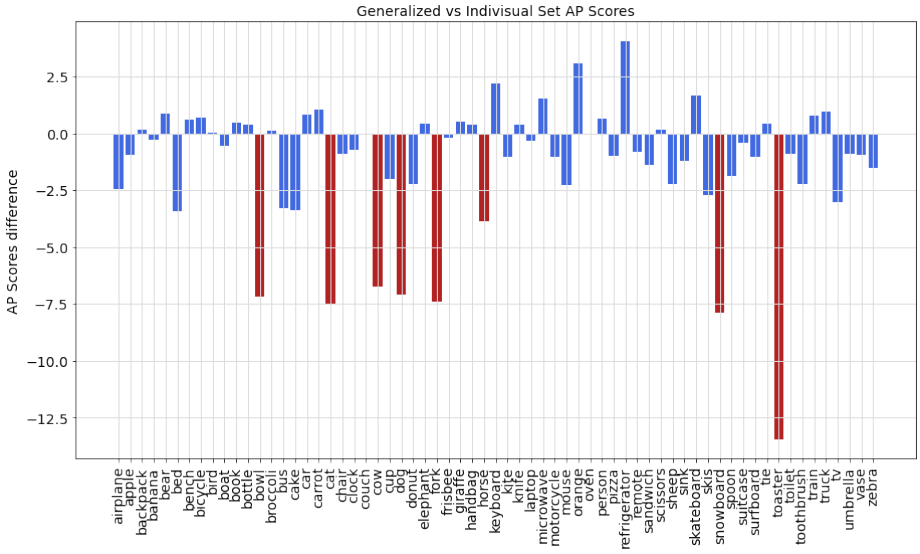


Fig. 2: We plot the difference in AP score when considering the generalized setup (all classes together) as compare to considering the individual sets of known and novel separately. Most of the classes present a drop when considering all classes together. Red bars correspond to classes with a drop larger than 3.5 AP.

couch and bed; or snowboard, skis, and skateboard or are sometimes confused by our model.

## References

1. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision (2017)
2. Pham, K., Kafle, K., Lin, Z., Ding, Z., Cohen, S., Tran, Q., Shrivastava, A.: Learning to predict visual attributes in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
3. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)

(a) Ground Truth (b) Our Results (c) Ground Truth (d) Our Results

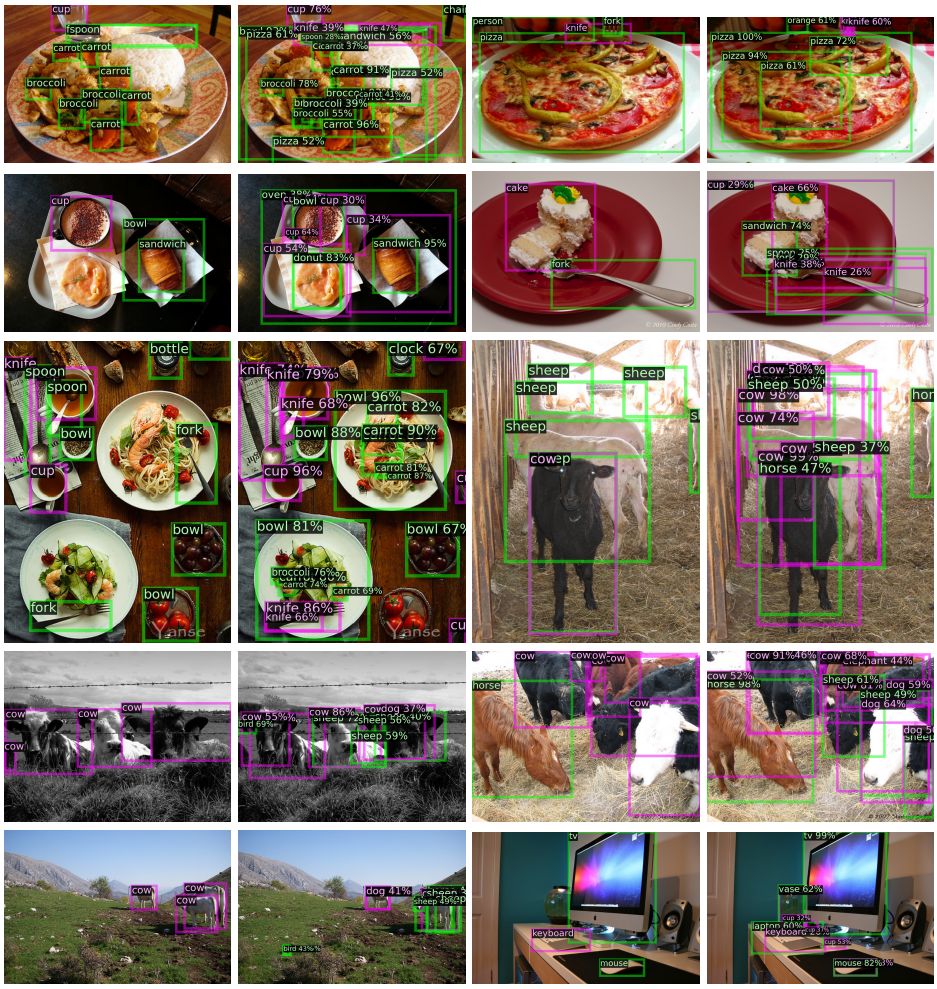

Fig. 3: Qualitative results obtained using LocOv on the COCO dataset. Novel classes are shown in magenta while known are in green. (Best viewed in color)

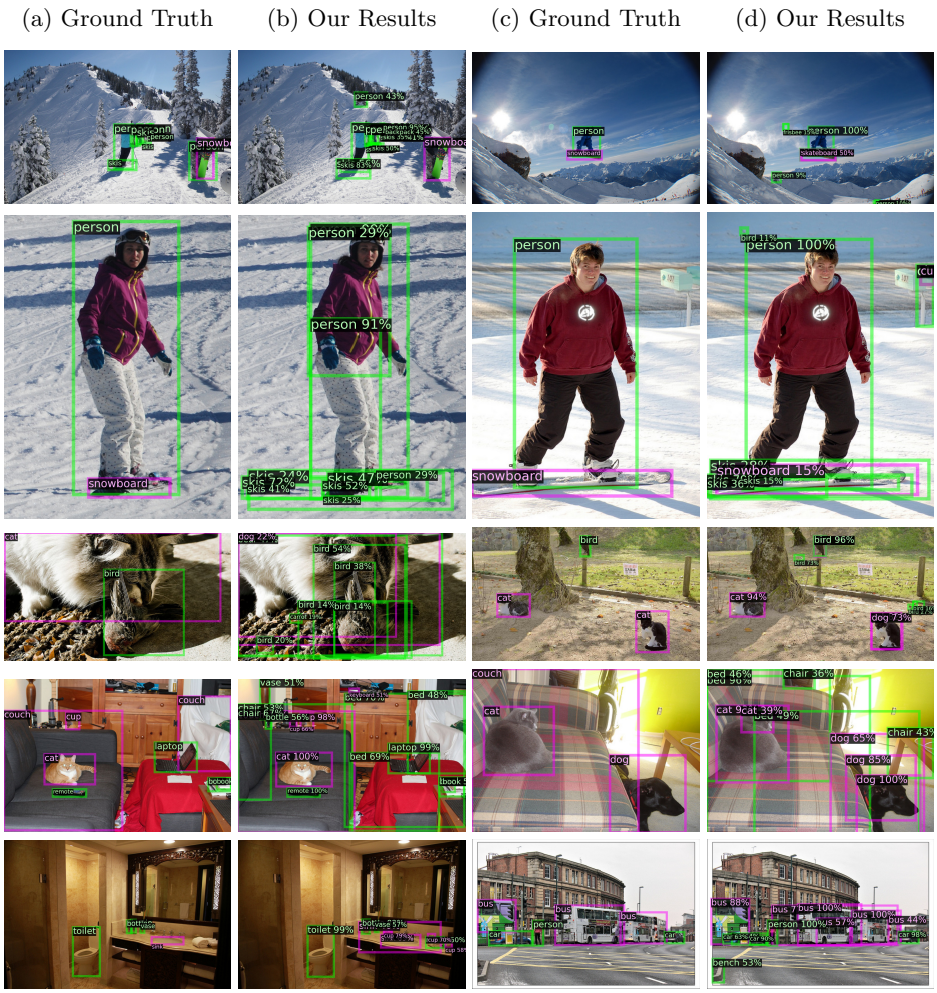


Fig. 4: Qualitative results obtained using LocOv on the COCO dataset. Novel classes are shown in magenta while known are in green. (Best viewed in color)