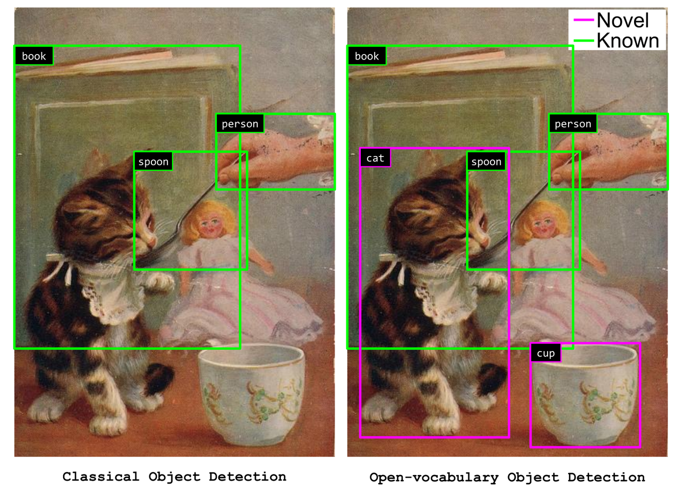
Localized Vision-Language Matching for Open-vocabulary Object Detection

María A. Bravo Sudhanshu Mittal Thomas Brox
{bravoma, mittal, brox}@cs.uni-freiburg.de

Motivation

- Traditional object detectors work in a **closed-world** setting → restrict detections to only discover **known** annotated objects. Scaling annotations to all possible objects in the world is infeasible.
- **Vision-language modeling** has made it possible for models to expand their vocabulary to **novel** objects without costly annotation.

Open-vocabulary Detection Task

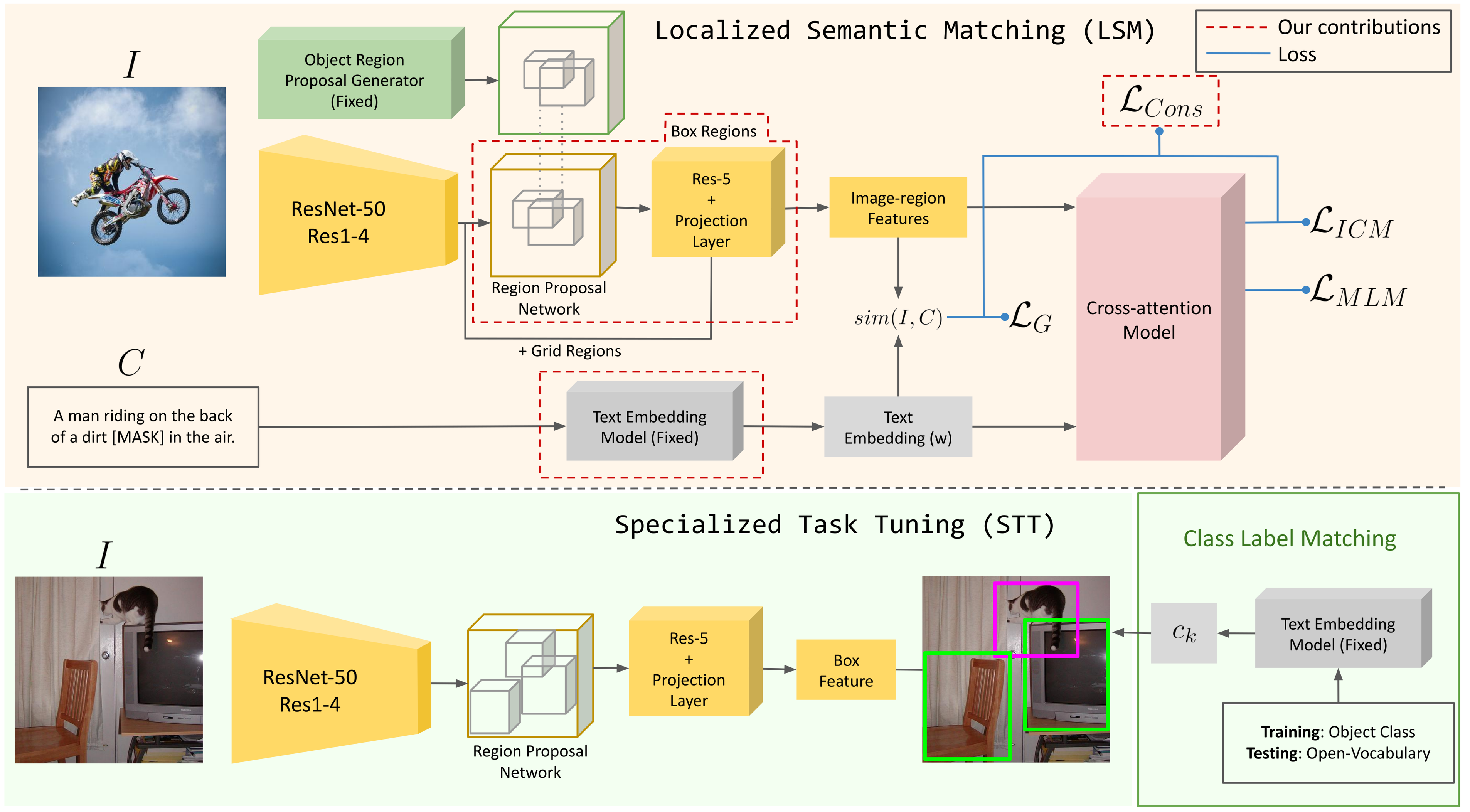


Objective: Train an object detector capable of detecting any object represented by a **text query**.
Training data types:

Object detection labels	Image-caption pairs
Bounding box and object class labels of known classes O_K	Images with corresponding caption description

Test object class sets:
O_K → known classes seen during training
O_N → novel classes no explicit annotations

Approach Overview



We propose **LocOv**, **Loc**alized Image-Caption Matching for **O**pen-**v**ocabulary, a two-stage approach with a Faster R-CNN [1] architecture.

1. *Localized Semantic Matching* (LSM): learn the semantics of objects in the image by matching image-regions to the words in the caption, learning a grounding loss $\mathcal{L}_G$. We exploit the multi-modal information by using a cross-attention model and an Image-Caption matching loss $\mathcal{L}_{ICM}$, the mask language modeling loss $\mathcal{L}_{MLM}$ and a consistency-regularization loss $\mathcal{L}_{Cons}$.
2. *Specialized Task Tuning* (STT): tunes the model using the known class annotations and specializes the model for object detection.

We define the sets:

$$R^I = \{r : r \text{ is an image-region feature vector from the image } I\}$$
$$W^C = \{w : w \text{ is a word or part-of-word feature vector from the caption } C\}$$

Method

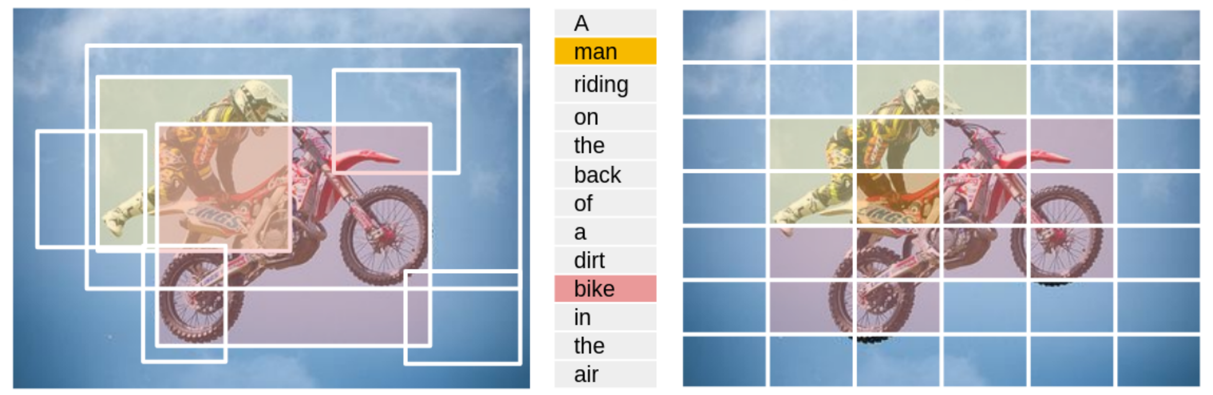
1. **Localized object region-text matching**
→ Match objects in the image to their corresponding class labels in the caption.
We define a contrastive **Grounding loss** based on the similarity between an image I and a caption C.

$$\mathcal{L}_{G_r}(I) = -\log \frac{\exp(sim(I,C))}{\sum_{C' \in \text{Batch}} \exp(sim(I,C'))},$$

$$sim(I,C) = \frac{1}{|R^I|} \sum_{i=1}^{|R^I|} \sum_{j=1}^{|W^C|} d_{i,j}(r_i \cdot w_j), \qquad d(r_i, w_j) = d_{i,j} = \frac{\exp(r_i \cdot w_j)}{\sum_{j'=1}^{|W^C|} \exp(r_i \cdot w_{j'})}.$$

Apply the grounding loss to two types of image-regions (r) features.

$$\mathcal{L}_G = \mathcal{L}_{G_{box}}(C) + \mathcal{L}_{G_{box}}(I) + \mathcal{L}_{G_{grid}}(C) + \mathcal{L}_{G_{grid}}(I)$$



2. **Disentangled text features**
→ Use embeddings of the pre-trained BERT model as text representations for image-text matching instead of contextualized text representations.

3. **Consistency-regularization**
→ Regularize the direct grounding loss with a cross-attention model. The cross-attention model takes the image-regions R^I and text embeddings W^C and calculates three losses.

- Image-caption matching loss $\mathcal{L}_{ICM}$,
- Masked Language Modeling loss $\mathcal{L}_{MLM}$, and
- Consistency-regularization loss $\mathcal{L}_{Cons}$. We use the Kullback-Leibler divergence loss to impose this consistency by comparing the matching distribution of the image-caption pairs before and after the cross-attention model.

Ablation Experiments

Datasets:
Training data: MS-COCO dataset [2] as the object detection dataset, and COCO captions [3] as the image-caption dataset.
Test data: MS-COCO dataset, results reported on 17 O_N, 48 O_K, and generalized ($O_K \bigcup O_N$).
Evaluation metrics: mean Average Precision (AP), and Average Precision using two fixed thresholds at 0.5 (AP_{50}) and 0.75 (AP_{75}).

$\mathcal{L}_{Cons}$	R^I_{box}	BERT Emb.	Novel (17) AP	AP_{50}	AP_{75}	Known (48) AP	AP_{50}	AP_{75}	Generalized AP	AP_{50}	AP_{75}
×	×	×	14.3	25.6	14.4	28.1	47.8	29.3	23.7	40.9	24.5
×	✓	✓	15.4	27.9	15.2	32.2	52.1	34.1	26.3	43.6	27.3
✓	×	✓	15.5	27.1	15.4	32.2	52.1	33.9	27.1	44.5	28.2
✓	✓	×	16.7	29.7	16.7	33.4	**53.5**	**35.5**	**28.2**	**45.9**	29.5
✓	✓	✓	**17.2**	**30.1**	**17.5**	**33.5**	53.4	**35.5**	28.1	45.7	**29.6**

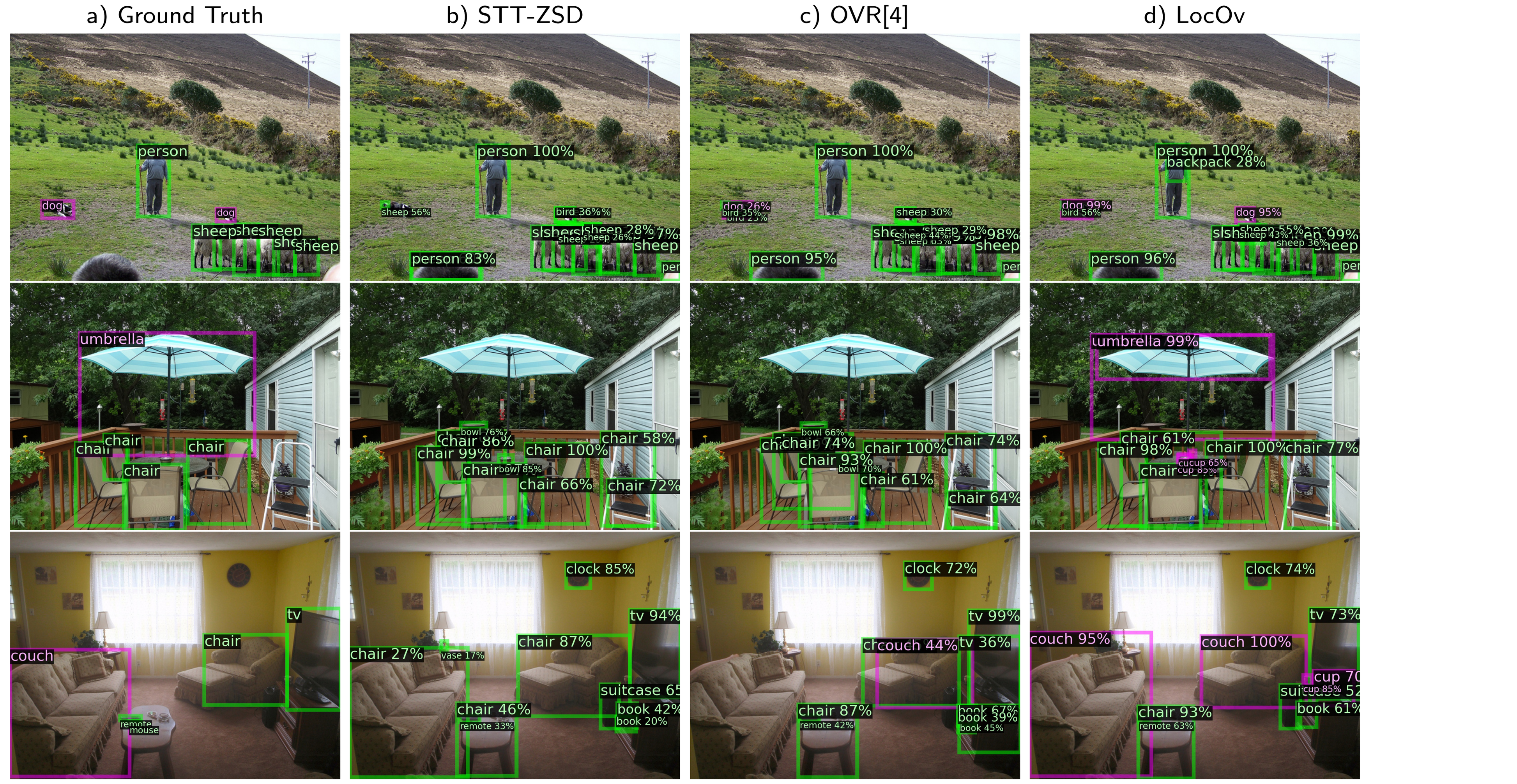
$\mathcal{L}_{Cons}$ = consistency-regularization, R^I_{box} = inclusion of box-regions together with grid-regions, BERT Emb. = BERT Embeddings only.

Results

Method	Img-Cap Data Size	Constrained Novel (17) AP	AP_{50}	Constrained Known (48) AP	AP_{50}	Generalized Novel (17) AP	AP_{50}	Generalized Known (48) AP	AP_{50}	All (65) AP	AP_{50}
Faster R-CNN	-		-	-	54.5	-	-	-	-	-	-
STT-ZSD (Ours)		0.21	0.31	33.2	53.4	0.03	0.05	**33.0**	53.1	24.4	39.2
OVR$^{*\S c}$ [4]	0.6M	14.6	27.5	26.9	46.8	-	22.8	-	46.0	22.8	39.9
LocOv $^{*\S c}$ (Ours)		**17.2**	30.1	**33.5**	53.4	**16.6**	**28.6**	31.9	51.3	**28.1**	45.7
XP-Mask$^{\ddagger\S\star c}$ [5]	5.7M	-	29.9	-	46.8	-	27.0	-	46.3	-	41.2
CLIP (cropped reg)† [6]	400M	-	-	-	-	-	26.3	-	28.3	-	27.8
RegionCLIP$^{\dagger\S c}$ [7]	400.6M	-	**30.8**	-	**55.2**	-	26.8	-	54.8	-	47.5
ViLD$^{\dagger c}$ [6]	400M	-	-	-	-	-	27.6	-	**59.5**	-	**51.3**

LocOv outperforms all other methods for Novel objects in the generalized setup while using only 0.6M of image-caption pairs.
Training dataset: *ImageNet1k, §COCO captions, †CLIP400M, ‡Conceptual Captions, *Open Images, and ^{c}COCO.

a) Ground Truth b) STT-ZSD c) OVR[4] d) LocOv



Conclusions

- The proposed localized matching technique helps in learning labels of novel classes as compared to only using grid features.
- Language embedding features are preferable over contextualized features for novel object detection.
- Consistency-regularization between grounding and cross-modal matching is crucial for the open-vocabulary detection task.

References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NeurIPS*, 2015.
[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.
[3] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.
[4] A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang, "Open-vocabulary object detection using captions," in *CVPR*, 2021.
[5] D. Huynh, J. Kuen, Z. Lin, J. Gu, and E. Elhamifar, "Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling," *arXiv preprint arXiv:2111.12698*, 2021.
[6] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," in *ICLR*, 2022. [Online]. Available: https://openreview.net/forum?id=lL3lnMbR4WU
[7] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li *et al.*, "Regionclip: Region-based language-image pretraining," *arXiv preprint arXiv:2112.09106*, 2021.

Acknowledgements

This work was supported by Deutscher Akademischer Austauschdienst - German Academic Exchange Service (DAAD) Research Grants - Doctoral Programmes in Germany, 2019/20; grant number: 57440921. The Deep Learning Cluster used in this work is partially funded by the German Research Foundation (DFG) - 417962828.