Betrayed by Captions: Joint Caption Grounding and Generation for Open Vocabulary Instance Segmentation

Jianzong Wu[1*] Xiangtai Li[2* †] Henghui Ding[2] Xia Li[3]
Guangliang Cheng[4] Yunhai Tong[1] Chen Change Loy[2]

[1] Key Laboratory of Machine Perception, MOE, School of Artificial Intelligence, Peking University
[2] S-Lab, Nanyang Technological University [3] ETH Zurich [4] SenseTime Research
jzwu@stu.pku.edu.cn {xiangtai.li, henghui.ding, ccloy}@ntu.edu.sg

A. More Implementation Details

Baseline Details. All the table results in main paper use **the same ResNet50 [4] backbone** for a fair comparison. The number of object queries is *100* by default for all experiments. Our method is trained by only 12 epochs on the COCO training set and evaluated on the COCO validation set. All the experiments are carried out on 8 V100 GPUs. Following previous methods [5, 9], we use mAP (mean AP on the IoU threshold of 0.5) as the metric for OVIS.

Training and Inference Details. We adopt the default training of Mask2Former [8, 2, 1]. A learning rate multiplier of 0.1 is applied to the backbone. For data augmentation, we use the default large-scale jittering (LSJ) augmentation with a random scale sampled from the range 0.1 to 2.0 with the crop size of 1024 $\times$ 1024. We use the default Mask R-CNN inference setting [3], where we resize an image with a shorter side to 800 and a longer side to 1333. *For the inference of OSPS*, we do not use the default joint merge for things and stuff that is used in Mask2Former [2]. Instead, we put the thing mask first and fill the remaining area with stuff mask prediction. In the experiment part, we find that the thing predictions for the unknown are usually in a low score, and they may be covered by high score stuff mask prediction. This is because all the stuff masks are trained in a supervised manner.

Training Splits For OVIS and OSPS. For OVIS, we follow the 48/17 split in COCO proposed by [7], in which 48 classes are base classes, and 17 are novel classes. For OSPS, we follow the unknown things split proposed by [6]. The unknown percentages are 5%, 10%, and 20% separately.

Concretely, for 48/17 split of OVIS, the **base** classes are: "person", "bicycle", "car", "motorcycle", "truck", "boat", "bench", "bird", "horse", "sheep", "zebra", "giraffe", "backpack", "handbag", "skis", "kite", "surfboard", "bottle", "spoon", "bowl", "banana", "apple", "orange", "broccoli", "carrot", "pizza", "donut", "chair", "bed", "tv", "laptop", "remote", "microwave", "oven", "refrigerator", "book", "clock", "vase", "toothbrush", "train", "bear", "suitcase", "frisbee", "fork", "sandwich", "toilet", "mouse", "toaster".

The **novel** classes are: 'bus', 'dog', 'cow', 'elephant', 'umbrella', 'tie', 'skateboard', 'cup', 'knife', 'cake', 'couch', 'keyboard', 'sink', 'scissors', 'airplane', 'cat', 'snowboard'.

For OSPS, the **unknown** things are: 5%: "car", "cow", "pizza", "toilet". 10%: "boat", "tie", "zebra", "stop sign". 20%: "dining table", "banana", "bicycle", "cake", "sink", "cat", "keyboard", "bear".

Table 1: OVR-CNN experiments for Open Vocabulary Object Detection on COCO with 48/17 spilt. "Vanilla" means the origin OVR-CNN model without our proposed modules.

Method	Constrained		Generalized		
	Base	Novel	Base	Novel	All
Vanilla	40.6	22.6	39.8	18.5	34.2
w. Gro	40.3	23.3	39.4	19.6	34.2
w. Gro & Gen	40.6	23.4	40.3	18.9	34.7

B. More Experiments Results

Will Joint Grounding and Captioning Help Other Architectures? We conduct experiments on a previous model, OVR-CNN [9], to further evaluate the effectiveness of our proposed modules, i.e., caption grounding with object nouns and caption generation. We re-implement OVR-CNN using PyTorch and add the two modules onto its architecture. The training schedule and results may differ from the original paper [9], while the training settings are the same in our experiments. Concretely, we train 40,000 steps with a batch size of 56 for the caption pre-training stage and

*The first two authors contributed equally to this work. † Corresponding Author and Leader. Code and model are available at https://github.com/jianzongwu/betrayed-by-captions.

Table 2: Ablation on fully supervised instance segmentation, object detection, and panoptic segmentation. AP-novel indicates the mean AP on the 17 novel classes (trained in the fully supervised setting). AP-bbox indicates object detection.

Method	Instance			Panoptic		
	AP	AP-novel	AP-bbox	PQ	PQ-th	PQ-st
class-label	59.3	66.6	58.9	46.4	51.9	38.2
class-emb.	50.6	57.8	50.2	44.4	50.5	35.1
w/ gro.	50.8	57.4	50.3	44.1	50.3	35.0
w/ gen.	50.9	57.6	50.7	44.2	50.5	34.8
w/ both.	51.3	57.5	50.7	44.3	50.6	34.9

Table 3: Ablation on layers of Caption Generator and quality of Open Vocabulary Instance Segmentation. We adopt BLUE, CIDEr, and ROUGE as the metrics to evaluate the quality of generated captions.

Layers	Segmentation			Caption Generation					
	Base	Novel	All	BLUE-1	BLUE-2	BLUE-3	BLUE-4	CIDEr	ROUGE
2	46.7	23.4	40.6	0.473	0.311	0.206	0.141	0.307	0.360
4	46.0	28.4	41.4	0.418	0.258	0.166	0.111	0.239	0.320
6	48.2	26.9	42.6	0.387	0.226	0.138	0.088	0.171	0.289

30,000 steps with a batch size of 48 for the detector fine-tuning stage. Tab. 1 shows that by adding caption grounding with object nouns, the novel AP score increases, indicating our proposed method's effectiveness. However, adding a caption generation module does not bring further improvement. This may be explained by the fact that OCR-CNN already applies ITM and MLM losses as auxiliary losses during the pre-training process, which extracts knowledge from all words in the captions.

Will Joint Grounding and Captioning Help the Fully Supervised Baseline? To answer this question, we perform ablation on fully supervised settings in Tab. 2. For the proposed CGG, we verify two main components: caption grounding and generation. Class-emb means only using pre-trained text embeddings for mask classification. Class-label is a traditional learnable, fully connected layer that converts the classes into contiguous labels. In Tab. 2, we observe that the fully supervised method achieves better results than using class embeddings in all three tasks. As shown in the last three rows of Tab. 2, for within-class embedding settings, the added caption grounding and generation modules help to improve the performance on OVIS but bring no performance gain on OSPS. We conclude that joint grounding and captioning have limited benefits (0.5% improvements) in supervised settings.

Will Better Caption Generator Help Open Vocabulary Instance Segmentation? We further explore the influence of the caption generation module to open vocabulary instance segmentation. Tab. 3 shows the results. As we adopt a larger caption generator, the overall segmentation quality (AP all) increases. On the contrary, the quality of the caption (including BLUE and CIDEr) generation drops. A better caption generator may not be a better open vocabulary instance segmenter. The role of the caption generator is to force the model to know the existence of novel objects, so pursuing a better caption generation model is not our goal for OVIS and OSPS.

Table 4: Ablation on different object nouns parsers.

Method	Novel AP	All AP
ImageNet 21K parser	19.9	41.3
LVIS parser (ours)	**28.4**	**41.4**

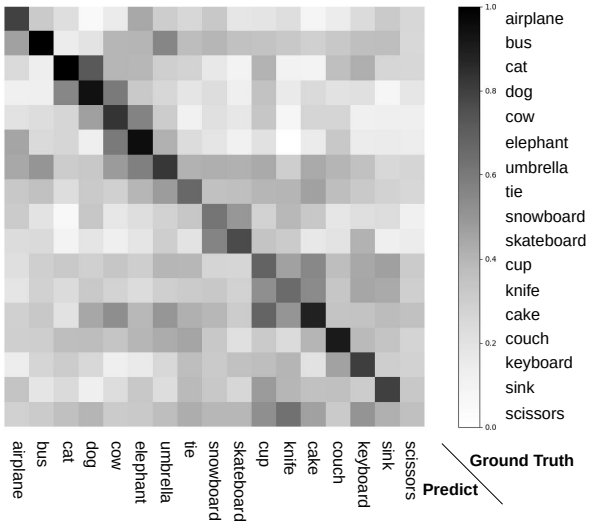

Figure 1: The correlation map between Ground Truth and model predictions on **novel classes**. The noun embeddings and object queries for novel classes are highly correlated.

ImageNet21K Parser to Extract Object Nouns We perform an ablation on the choice of object noun parser. The results are shown in Tab. 4. When using ImageNet 21K parser, the novel AP drops from 28.4 to 19.9. This could potentially stem from the fact that many class names within ImageNet 21K comprise words that are not object nouns, such as "drive", "yellow", "red", and "top".

C. Visual Analysis and Comparison

Visualization Analysis both Nouns and Object Queries. We calculate the correlation map between the predicted multi-modal embeddings e_i and the Ground Truth class embeddings. As shown in Fig. 1, our model can correctly distinguish novel classes based on the segmentation masks.

More Visual Examples from Caption Generation. We observe that in some cases, the caption generated by CGG

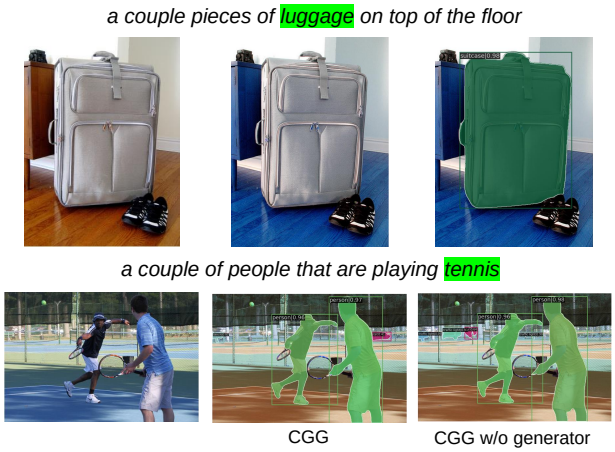

Figure 2: Examples of captions predicting objects that are not in the category list.

can predict objects that are *not* in the category list. Categories beyond the given list cannot be correctly classified using the similarity between multi-modal embeddings and class embeddings since the class embeddings are not accessible during inference, like in top images of Fig. 2. There is a couple of luggage on the floor, but "luggage" is not a class in the validation dataset. *Without* a caption generator, the model classifies the luggage as "suitcase". However, with the caption generation module, the generated caption successfully depicts the word "luggage". In the bottom images, "tennis" is also described by captions. Fig. 3 shows more visualization results with captions.

More Visualization Results on OVIS and OSPS. In Fig. 4, we present more visual results of OVIS and OSPS tasks. The CGG model can well segment and classify novel categories well.

Zero Shot Visualization on ADE20K dataset. In Fig. 5, we show the visualization results on ADE20K dataset [10]. CGG can detect and segment novel classes in a zero-shot manner on ADE20K. At the same time, CGG generates comprehensive captions that well depict the content of the images.

References

[1] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 1

[2] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. *CVPR*, 2022. 1, 5

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1

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

[5] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *CVPR*, 2022. 1

[6] Jaedong Hwang, Seoung Wug Oh, Joon-Young Lee, and Bohyung Han. Exemplar-based open-set panoptic segmentation network. In *CVPR*, 2021. 1

[7] Shafin Rahman, Salman Khan, and Fatih Porikli. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In *ACCV*, 2018. 1

[8] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 1

[9] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. *CVPR*, 2021. 1

[10] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 3, 5

Figure 3: Visualization results of generated captions and the related segmentations of CGG. Input Image (Left), CGG w/o caption generation (Middle), CGG (Right). "mirror" and "fire hydrant" are not in the category list (both base and novel) but are still mentioned in the generated captions.

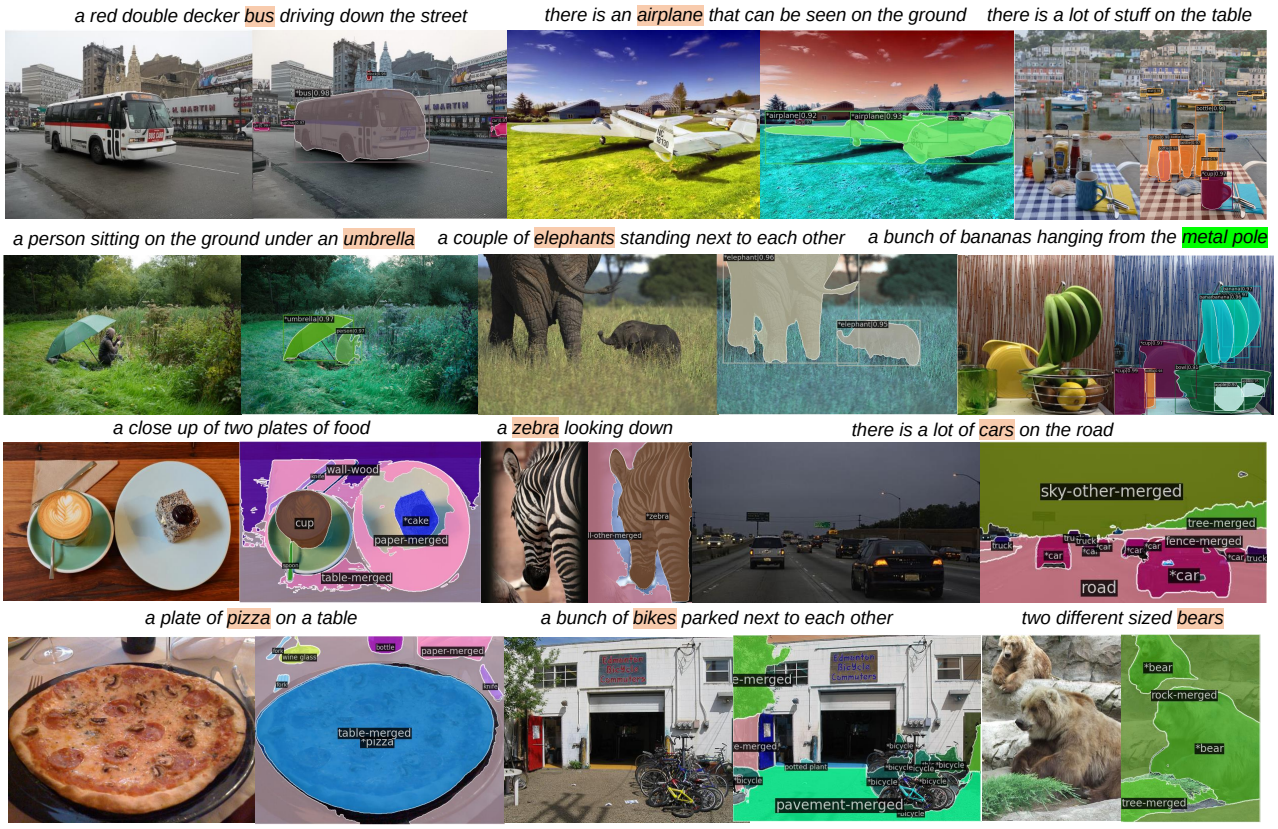

Figure 4: More visualization results of OVIS (Top two rows) and OSPS (Bottom two rows). Novel classes are marked by "*".

Figure 5: Visualization on ADE20k [10]. Following [2], we apply instance segmentation on 100 instance classes. Classes not in COCO are marked by "*".