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

Abstract

In this work, we focus on instance-level open vocabulary segmentation, intending to expand a segmenter for instance-wise novel categories ***without*** *mask annotations. We investigate a simple yet effective framework with the help of image captions, focusing on exploiting thousands of object nouns in captions to discover instances of novel classes. Rather than adopting pretrained caption models or using massive caption datasets with complex pipelines, we propose an end-to-end solution from two aspects: caption grounding and caption generation. In particular, we devise a joint* ***Caption Grounding and Generation (CGG)*** *framework based on a Mask Transformer baseline. The framework has a novel grounding loss that performs explicit and implicit multi-modal feature alignments. We further design a lightweight caption generation head to allow for additional caption supervision. We find that grounding and generation complement each other, significantly enhancing the segmentation performance for novel categories. We conduct extensive experiments on the COCO dataset with two settings: Open Vocabulary Instance Segmentation (OVIS) and Open Set Panoptic Segmentation (OSPS). The results demonstrate the superiority of our CGG framework over previous OVIS methods, achieving a large improvement of* ***6.8% mAP*** *on novel classes without extra caption data. Our method also achieves over* ***15% PQ*** *improvements for novel classes on the OSPS benchmark under various settings.*

1. Introduction

Instance-Level Segmentation [17, 40] is a core vision task that goes beyond object detection [38, 39, 50] via segmenting and classifying each object. Although it continues to attract significant research effort [2, 4–6, 8–10, 18, 26, 34–37, 44, 54–57, 60, 61, 64, 72, 74], current solutions mainly focus on a closed-set problem that assumes a *pre-defined set of object categories* [24, 31, 40]. In practice, many applications need to detect and segment new categories. To save the need of annotating new object categories, zero-shot object detection/segmentation [3, 48] is proposed, where models are trained on base classes and equipped with the ability to segment new classes. However, zero-shot setting suffers from low novel-class performance, as high-level word embeddings cannot effectively encode fine-grained visual information.

To address this issue, recent work [69] takes an open vocabulary setting by pretraining a visual backbone on captioned images for learning rich visual features. With the success of pretrained Vision Language Models (VLMs) [30, 46], several approaches, *e.g.*, ViLD [23], propose effective methods to distill knowledge from VLMs into detectors or segmentation methods. Meanwhile, several works decouple the learning of open vocabulary classification and detection/segmentation into a two-stage pipeline [15, 21]. Recently, state-of-the-art solutions [19, 28, 33, 71, 79] for open vocabulary detection/segmentation try to adopt larger-scale dataset pre-training with the help of VLMs. For example, Detic [79] adopts the ImageNet-21k [51] dataset to enlarge the detector in a weakly supervised manner, while PromptDet [19] augments the detection dataset with image-caption pairs scraped from the Internet. Recent XPM [28] also pretrains their model on caption datasets [52]. These approaches typically require a complex architecture design to leverage extra datasets [31, 51]. Despite the performance improvement, these methods are not cost-effective in terms of data utilization. In this paper, we explore the use of caption data with more effective designs.

Caption-related vision tasks can be broadly divided into grounding and generation. The former [13, 14, 22, 41, 67]

*The first two authors contribute equally to this work. Parts of the work are done when Xiangtai was an intern at SenseTime Research. † Project Leader. Code and model will be made available at https://github.com/jzwu48033552/betrayed-by-captions.

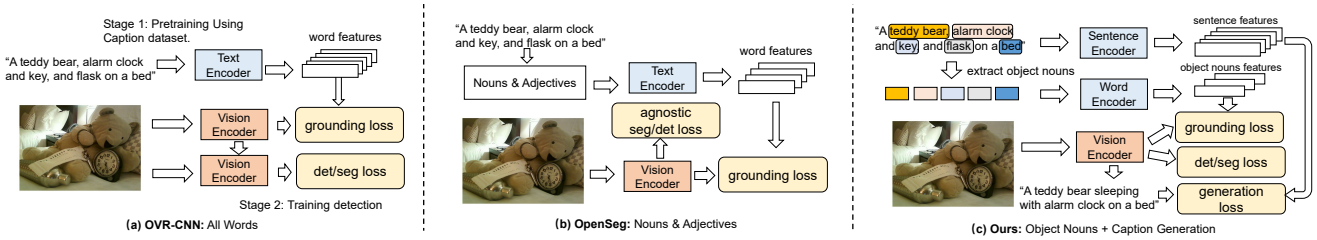

Figure 1. (a) OVR-CNN [69] uses all words for caption grounding, then finetunes, in a two-stage pipeline. (b) OpenSeg [21] uses extra agnostic head for segmentation and Nouns for grounding. (c) Our method encodes only object nouns in captions for caption grounding, and all words for caption generation in one unified framework.

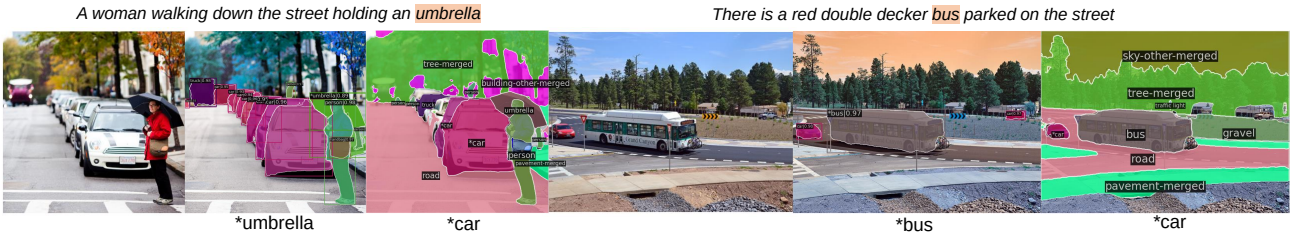

Figure 2. Example of Instance Segmentation and Panoptic Segmentation results of CGG. Categories marked by '*' are novel categories. Sentences are generated by the Instance Segmentation model. To the best of our knowledge, we are the first to unify OVIS, OSPS and Caption Generation in one framework.

requires a model to align the text and corresponding region features, *e.g.*, OVR-CNN [69] and OpenSeg [21] in Fig. 1 (a) and (b), while the latter [59, 66, 73] learns a model that output a caption for a given imagery input. The relationship between the two tasks and open vocabulary instance segmentation is not well explored. We argue that caption data encode rich structural and semantic information, which may help the process of novel class detection. Different from the OVR-CNN [69] (see Fig. 1(a)) that adopts a caption model in the pre-training process where the caption data and detection results are not well aligned, we propose a unified framework to *jointly* perform caption grounding, generation, and instance segmentation.

Our framework presents a novel caption grounding loss and an extra caption decoder for the generation loss, as shown in Fig. 1(c). The caption data is thus well exploited in both the input and output stages. In particular, we use object queries as inputs following Mask2Former [10]. At the input stage, we adopt separated object nouns to ground each object query, providing us with the grounding loss. At the output stage, with a lightweight Transformer decoder, we add supervision to the generated caption, resulting in the generation loss. Both losses are well coupled and have a mutual effect for novel class segmentation, adding **only 0.8%** GFlops during the training. For inference, our method drops the caption generation module for OVIS and OSPS with no extra computation cost.

We carry out experiments in two different settings, including Open Vocabulary Instance Segmentation (OVIS) and Open Set Panoptic Segmentation (OSPS) [29]. Experimental results demonstrate that our proposed method achieves significant improvements for novel classes despite using a strong baseline [10] as the encoder. The proposed method achieves new state-of-the-art results on COCO OVIS and COCO OSPS *without any data pre-training and complex pipelines*. Figure 2 shows that our method predicts instance segmentation, panoptic segmentation, and the corresponding caption in *one unified* framework while predicting novel classes. In particular, our method achieves a large improvement of **6.8% mAP** over previous XPM [28] on OVIS and **15% PQ** improvements over previous method [63].

2. Related Work

Zero-Shot Detection and Segmentation. Scaling up data collection and annotation is laborious and expensive for large vocabulary detection and segmentation. Zero-Shot Detection [48] and Segmentation [3] tries to detect/segment novel categories that the annotations are not accessible during the training process. Many studies address this problem by aligning region features to the fixed text embeddings [1, 20, 47, 70, 80]. Due to the limited capacity of word embeddings and the advent of large Vision-Language-Models (VLMs), recent studies [23, 68, 69] have moved to the open vocabulary setting.

Open Vocabulary Object Detection (OVOD). Recent studies [16, 23, 68, 69, 79] focus on the open vocabulary set-

ting, in which models are trained by leveraging pre-trained language-text pairs including captions and text prompts. For instance, OVR-CNN [69] is first pretrained on image-caption data to recognize novel objects, then fine-tunes the model for zero-shot detection. Recently, many works on image classification successfully expand their vocabulary sizes by pretraining on large-scale image-text pairs datasets. ViLD [23] proposes to distill the rich representation of pre-trained CLIP [46] into the detector, while DetPro [16] adds a fine-grained automatic prompt learning. Meanwhile, several works extract pseudo region annotations from the pre-trained VLMs and employ them as the additional training data for detectors. Detic [79] improves the performance on the novel classes with image classification datasets by supervising the max-size proposal with all image labels. Methods above share the same idea of trying to enlarge the capacity of training data to find the rare classes, thus they need more computation/annotation costs and complex pipelines. On the contrary, we focus on designing a way to discover novel classes from the caption data in one unified framework ***without*** pre-training on extra datasets nor distilling knowledge from pretrained VLMs.

Open Vocabulary Segmentation (OVS). Beyond OVOD, OVS further requires the model to segment the novel classes. Current solutions for OVS usually decouple mask generation and mask classification as two different steps. The former generates mask regions, while the latter performs classification with pre-trained VLMs [21,32]. DenseCLIP [78] proposes a similar pipeline to that in OVD by distilling CLIP knowledge through generating pseudo mask labels. Our method proposes an end-to-end pipeline to perform caption learning (grounding/generation) and segmentation learning jointly. The differences with OpenSeg [21] are: 1. We extract object nouns from captions, rather than nouns and adjectives as in OpenSeg. 2. For text encoders, we use BERT embeddings that are purely trained on text corpus, while OpenSeg employs a state-of-the-art VLM (ALIGN [30]). 3. We mainly focus on instance-level open vocabulary segmentation task rather than semantic segmentation.

Image Captioning. This task requires the model to generate captions to describe the content of images [59]. State-of-the-art methods follow multi-modal attention designs, treating the task as a multi-modal translation problem [66, 73, 75]. Our focus in this work is not to design a new captioning model, but to explore image captioning as a sub-task for open vocabulary learning to enhance the novel class discovery ability. To our best knowledge, this study is the first attempt that explores caption generation on OVS.

3. Methodology

In this section, we first review the background of open vocabulary instance segmentation. Then, we present our Caption Grounding and Generation framework, which aims to fully exploit caption data through joint caption grounding and generation.

3.1. Background

Problem Setting. We first describe the open-vocabulary problem setting. Let $\mathcal{D}_B = \{(\boldsymbol{I}_m, \mathcal{Y}_m)\}_{m=1}^{N_B}$ be the set of training images and instance annotations for a limited set of base classes $\mathcal{V}_B$. Among these images, there are also novel classes $\mathcal{V}_N$, whose annotations cannot be accessed during the training. For OSPS, novel classes come from the thing classes, while the stuff classes are treated as base classes. Each image $\boldsymbol{I}_m$ is associated with a set of ground-truth (GT) annotations $\mathcal{Y}_m$, which comprises instance masks and their corresponding object classes. In order to detect and segment novel classes, following previous works [69], we leverage additional image-level annotations, i.e., image captions. Let $\mathcal{D}_C = \{(\boldsymbol{I}_c, \mathcal{Y}_c)\}_{c=1}^{N_C}$ be another set of training images with image caption annotations. Each image $\boldsymbol{I}_c$ is annotated with a caption $\mathcal{Y}_c$. Compared to pixel-level annotations, captions are easier to collect, and its vocabulary $\mathcal{V}_C$ is much larger than base classes, i.e., $|\mathcal{V}_C| \gg |\mathcal{V}_B|$. Therefore, exploiting the additional information from the image caption dataset would be beneficial.

Open-vocabulary instance segmentation aims to train a model to segment both base classes $\mathcal{V}_B$ and novel classes $\mathcal{V}_N$. Following previous methods [21, 28, 69], our model uses high-level semantic embeddings from a pretrained text Transformer (BERT [12]) as the weights of the linear classifier. We focus on distilling knowledge in the captions to the target classes via representation similarities.

Baseline Method. We adopt the recent Mask2Former [10] model as our baseline since the mask-based Transformer architecture can be readily extended into multi-modal training with captions. Mask2Former takes a Transformer encoder-decoder architecture with a set of object queries, where the object queries interact with encoder features via masked cross-attention. Given an image I, during the inference, Mask2Former directly outputs a set of object queries $\boldsymbol{Q} = \{q_i\}, i = 1, ..N_q$, where each object query q_i represents one entity. Then, two different Multiple Layer Perceptrons (MLPs) project the queries into two embeddings for mask classification and mask prediction, respectively. During the training, each object query is matched to the ground truth mask via masked-based bipartite matching. The loss function is $L_{mask} = \lambda_{cls} L_{cls} + \lambda_{ce} L_{ce} + \lambda_{dice} L_{dice}$, where L_{cls} is the Cross-Entropy (CE) loss for mask classification, and L_{ce} and L_{dice} are the Cross-Entropy (CE) loss and Dice loss [43] for segmentation, respectively. In particular, following [69], we use pretrained embeddings to replace the learnable classifier for training and inference, as shown in Fig. 3.

The original Mask2Former can only detect and segment

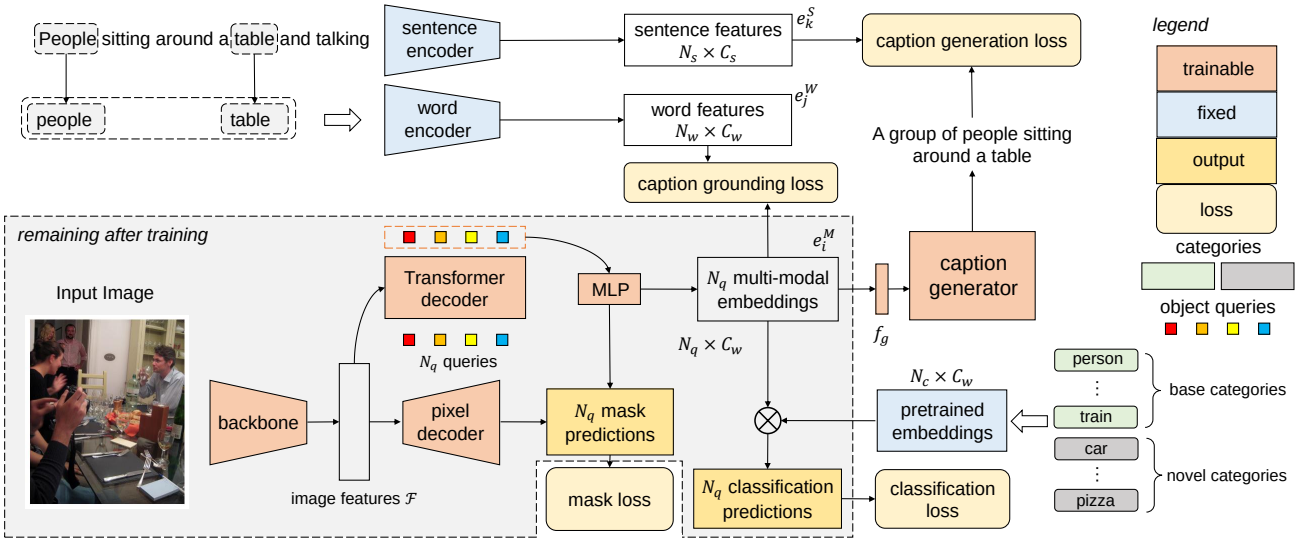

Figure 3. The illustration of CGG framework. The input image I is first provided to Mask2Former. The output of the Transformer decoder is then fed into an MLP, which generates N_q mask predictions together with the output of the pixel decoder. Then the query is transferred into N_q multi-modal embeddings $\{e_i^M\}$, of which the similarity with class embeddings is computed to produce classification predictions. $\{e_i^M\}$ also outputs grounding loss and generation loss with text features extracted by word encoder and sentence encoder.

closed-set objects and cannot handle the novel classes. Our method extends it to perform open-vocabulary segmentation in a new framework.

3.2. CGG Framework for OVS

Overview. Figure 3 presents the overall pipeline of our CGG framework. Based on Mask2Former [10], following [69], we set the pretrained text embeddings as the weights of the final linear classifier. We add two losses: the caption grounding loss and the caption generation loss. A caption generator is appended at the end of the output queries, directly producing the image caption. During the training, we adopt a pre-trained sentence encoder and word encoder to encode both captions and object nouns extracted from captions into sentence features and word features. The former is used for caption generation loss, while the latter is used for caption grounding loss. During the inference, we discard all the newly-introduced modules, and perform the same inference procedure as Mask2Former.

Class-Agnostic Pretraining. Following previous works [21, 28, 69], we first pretrain our framework using *only* base data annotations in a class-agnostic manner. Such a process is similar to training a Region Proposal Network (RPN) at the first stage. The goal of pretraining is to encode instance-wised information into object queries. Then we load the pretrained model for joint training with caption data.

End-to-End Caption Grounding. Previous works like OVR-CNN [69] pre-train their models with caption data. The core idea is to learn a Vision to Language (V2L) projection layer where the language data from novel classes are transformed into vision features via multiple multi-modal losses including grounding loss and a set of auxiliary self-supervision loss.

There are two potential issues with the previous design. Firstly, training caption and segmentation separately cannot fully explore caption data and detection/segmentation annotations. The training of segmenter is isolated and the connection between the two models is broken. Secondly, there is a weakened region-word alignment in the traditional grounding process by calculating similarities between multi-modal embeddings and **all** words in caption data, because object-unrelated words may encounter the vision-language implicit matching. We argue that object nouns in caption data should be well aligned with region features in a more fine-grained manner since the class categories are always nouns in captions.

Therefore, rather than sending the entire sentence as inputs, we extract only object nouns from the sentence and feed it to a word encoder. Such extraction finds more precise semantic embeddings, verified effectively for novel class grounding. For multi-modal embeddings, we adopt an MLP as the V2L projection and take the outputs of Transformer decoder in Mask2Former as the inputs, since the object queries group region features naturally, which is well proved in many previous works [5, 10]. Given an image-caption pair (I, C), we first calculate similarities between N_q multi-modal embeddings $\{e_i^M\}$ and N_w word features

$\{e_j^W\}$ extracted from the caption.

$$S_C(I,C) = \frac{1}{N_w}\sum_{j=1}^{N_w}\sum_{i=1}^{N_q} a_{i,j}^C \langle e_i^M, e_j^W \rangle,$$
$$S_I(I,C) = \frac{1}{N_q}\sum_{i=1}^{N_q}\sum_{j=1}^{N_w} a_{i,j}^I \langle e_i^M, e_j^W \rangle, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ is the dot product between two vectors. $S_C(I,C)$ and $S_I(I,C)$ are similarities between image I and caption C normalized for text or image separately. The normalization term is formulated as

$$a_{i,j}^C = \frac{\exp\langle e_i^M, e_j^M \rangle}{\sum_{i'=1}^{N_q} \exp\langle e_{i'}^M, e_j^W \rangle},$$
$$a_{i,j}^I = \frac{\exp\langle e_i^M, e_j^M \rangle}{\sum_{j'=1}^{N_w} \exp\langle e_i^M, e_{j'}^W \rangle}. \tag{2}$$

The core idea of grounding is that: the similarity between the matched image-caption pairs should be high, while for the unmatched pairs, it should be low. During the training, given a batch of image-caption pairs (B_I, B_C), for each similarity $S \in \{S_C, S_I\}$, the grounding loss is composed of two aspects. Take the similarity normalized along text dimension S_C as an example, from the image perspective, the grounding loss is as follows:

$$L_{gro}^{IC}(I) = -\log \frac{\exp S_C(I,C)}{\sum_{C' \in B_C} \exp S_C(I,C')}, \tag{3}$$

and from the caption perspective, the grounding loss is formulated as:

$$L_{gro}^{CC}(C) = -\log \frac{\exp S_C(I,C)}{\sum_{I' \in B_I} \exp S_C(I',C)}. \tag{4}$$

The final grounding loss is formulated as the sum of four losses:

$$L_{gro} = \frac{1}{|B_I|}\sum_{i=1}^{B_I}(L_{Gro}^{II}(I_i) + L_{Gro}^{IC}(I_i)) + \frac{1}{|B_C|}\sum_{j=1}^{B_C}(L_{Gro}^{CI}(C_j) + L_{Gro}^{CC}(C_j)). \tag{5}$$

Optimizing the grounding loss aligns the multi-modal embeddings and language embeddings in a large noun vocabulary.

End-to-End Caption Generation. Besides using caption data for grounding loss to align regions and words, we argue that caption data can also be employed as a generative supervision signal for a more fine-grained multi-modal understanding. The key insight is that we force the model to predict the occurring instances and their relationships in the image to identify novel classes. Unlike grounding loss that aims to push nouns and query embeddings as close as possible, generative loss decodes the visual features into the semantic embeddings, which are complementary to the grounding loss. As shown in Fig. 4, the caption generation module can help the model learn specific status and relationships of objects in the scene.

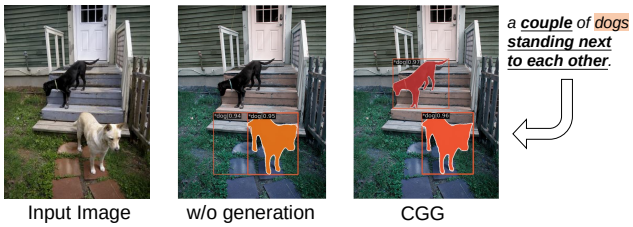

Figure 4. The effectiveness of caption generation. The generated caption depicts rich information beyond object nouns.

Specifically, since the multi-modal embeddings encode the region-wise information, we directly take these embeddings $\{e_i^M\}$ as the input of a lightweight caption generator, which includes a stack of Transformer decoder layers. To supervise the caption generator, we simply adopt a Cross Entropy Loss on the predicted distribution of text vocabularies, which is the commonly used objective function in the research field of caption generation.

$$L_{gen} = -\sum_{t=1}^{N_s} \log(p_\theta(w_t | w_1, ..., w_{t-1})), \tag{6}$$

where $p_\theta(\hat{w}_t | \hat{w}_1, \ldots, \hat{w}_{t-1})$ is the probability of predicting a particular word from the caption, θ denotes the parameters of the generation network. Hence, this loss function enforces the predicted sentence to be consistent with the input caption C, making the multi-modal embeddings $\{e_i^M\}$ capable of representing various objects and their potential relations in the image.

Overall Loss Design. The overall training loss contains four items, i.e., the classification loss L_{cls}, the segmentation loss L_{mask}, the caption grounding loss L_{gro}, and the caption generation loss L_{gen}. Following the previous method [69], the classification loss is selected as the Cross-Entropy Loss that takes the dot product of multi-modal embeddings e_i^M and base class embeddings as its logit inputs. The final loss function L is the weighted summation of the four losses: $L = \lambda_{cls}L_{cls} + \lambda_{mask}L_{mask} + \lambda_{gro}L_{gro} + \lambda_{gen}L_{gen}$. We follow the default setting in the MMDetection framework, where the weights are set to 2.0, 5.0, 2.0 and 2.0 in all our experiments.

Inference. Compared to the baseline model, CGG only introduces extra losses and a caption generation head during the training. During the inference, following [69], we use

the pretrained embeddings of all classes to perform open vocabulary segmentation via dot product, including base classes and novel classes. The remaining inference procedure is the same as the Mask2Former [10].

4. Experiments

4.1. Experimental Setup

Dataset Settings. For Open Vocabulary Instance Segmentation, we mainly conduct experiments on COCO dataset [40]. Following previous works [28, 69], we split 48 base classes with mask annotations and 17 target classes without mask annotations. There are 107,761 training images with 665,387 mask annotations from base classes and 4,836 testing images consisting of 28,538 and 4,614 mask instances for base and target classes, respectively. For captioned images, we use the entire MS-COCO training set with 118,287 images. Each image is annotated with five captions describing the visually-grounded objects in the image. Unlike previous works [19, 49, 79] that adopt extra caption datasets, like Conceptual Captions [53] with 3M image-caption pairs for pre-training, we do not use extra caption datasets or detection datasets. We follow the origin OVR-CNN [69] setting by only exploring a limited caption dataset within COCO. For Open Set Panoptic Segmentation [29], we mainly adopt the COCO-panoptic dataset. We follow the previous works [29, 63] by splitting part of thing classes into unknown classes. We obtain three different splits by varying the numbers of unknown classes ($K\%$ ratios, 5%, 10%, 20%). Different from previous works, we use extra caption data for training.

Metric. *For OVIS setting*, we report the mask-based mean Average Precision (mAP) at intersection-over-union (IoU) of 0.5. To analyze the performances on base and target classes, we carry out experiments in two settings: constrained setting where the model is only evaluated on test image inputs, which belong to either base classes or target classes; generalized setting in which a model is tested on both base and target class images. The latter is more challenging as it requires the model to segment target classes and avoid class bias from base classes, mostly with very high scores. We further report open vocabulary detection with box-based mAP. *For OSPS setting*, we employ the panoptic segmentation metrics, including Panoptic Quality (PQ) and Segmentation Quality (SQ), where we report known classes and unknown classes separately for reference. More details about the data preparation can be found in the appendix.

Implementation Details. We implement our models in PyTorch [45] with MMDetection framework [7]. For both settings, we use the distributed training framework with 8 GPUs. Each mini-batch has one image per GPU. The optimizer is AdamW [42] with a weight decay of 0.0001. We adopt full image size for a random crop in both the pretraining and training process following Mask2Former [10]. For classification head, word encoder, and sentence encoder, we all adopt the BERT embeddings (pre-trained with fixed input embeddings, not the output of transformer layers). We use a LVIS class name parser to extract object nouns from captions to ensure that the extracted nouns represent objects in the image [24]. For OVIS, we keep the top-100 queries as the model outputs. For OSPS, following previous work [29, 63], rather than Mask2Former baseline, we put thing mask predictions first and fill the remaining background with stuff mask predictions. ***Note that all experiments use ResNet-50 backbone for fair comparison.***

Table 1. Results on Open Vocabulary Instance Segmentation.

Method	Constrained		Generalized		
	Base	Novel	Base	Novel	All
OVR [69]	42.0	20.9	41.6	17.1	35.2
SB [1]	41.6	20.8	41.0	16.0	34.5
BA-RPN [76]	41.8	20.1	41.3	15.4	34.5
XPM [28]	42.4	24.0	41.5	21.6	36.3
CGG (Ours)	**46.8**	**29.5**	**46.0**	**28.4**	**41.4**

Table 2. Results on Open Set Panoptic Segmentation (OSPS). Note that previous methods EOPSN [29] and Dual [63] treat all unknown things as one class while not classifying them. In contrast, CGG performs complete Open Vocabulary Panoptic Segmentation, classifying each unknown thing into its specific category. We report the mean PQ and SQ for all unknown categories. * indicates that the scores are averaged from each unknown class.

Method	$K(\%)$	Known				Unknown	
		PQ^{Th}	SQ^{Th}	PQ^{St}	SQ^{St}	PQ^{Th}	SQ^{Th}
EOPSN [29]	5	44.8	80.5	28.3	73.1	23.1	74.7
Dual [63]		45.1	80.9	28.1	73.1	30.2	80.0
CGG (Ours)		**50.2**	**83.1**	**34.3**	**81.5**	**45.0***	**85.2***
EOPSN	10	44.5	80.6	28.4	71.8	17.9	76.8
Dual		45.0	80.7	27.8	72.2	24.5	79.9
CGG (Ours)		**49.2**	**82.8**	**34.6**	**81.2**	**41.6***	**82.6***
EOPSN	20	45.0	80.3	28.2	71.2	11.3	73.8
Dual		45.0	80.6	27.6	70.1	21.4	**79.1**
CGG (Ours)		**48.4**	**82.3**	**34.4**	**81.1**	**36.5***	78.0*

4.2. Main Results

Results on OVIS. We evaluate the performance of CGG and baselines on the Open Vocabulary Instance Segmentation task on MSCOCO. As shown in Tab. 1, our model outperforms the best baseline XPM by 5.5% mAP in the constrained setting where only novel categories input and 6.8% mAP in the generalized setting where both base and novel categories are employed as input. The generalized setting is more challenging because the model also needs to distinguish novel categories from given base categories, which have a data distribution bias from training data. We observed that CGG has a greater improvement in general-

Table 3. Results on COCO Open Vocabulary Object Detection (OVOD). IN-21K indicates ImageNet-21K [11]. CC indicates Conceptual Captions [53]

Method	Epochs	Extra Data	$AP50^{box}_{novel}$	$AP50^{box}_{all}$
DLWL [49]	96	YFCC100M	19.6	42.9
Cap2Det [65]	8.5	None	20.3	20.1
OVR-CNN [69]	12	None	22.8	39.9
Detic [79]	96	IN-21K & CC	24.1	44.7
PromptDet [19]	24	LAION-novel	26.6	50.6
CGG (Ours)	12	None	**29.3**	42.8

ized settings compared to constrained settings, which further shows the effectiveness of CGG in identifying the language embeddings of novel classes and distinguishing them from base classes.

Results on OSPS. To show the scalability of CGG, we also perform experiments on the Open Set Panoptic Segmentation task by expanding the base classes from base thing classes to including stuff classes, maintaining the whole training pipeline of CGG unchanged. The results are shown in Tab. 2. Compared with traditional Open Set Panoptic Segmentation, CGG actually performs a more difficult Open Vocabulary Panoptic Segmentation and still outperforms previous methods EOPSN and Dual by a large margin of 14.9% PQ on unknown things in 20% unknown things setting [29], and 16.9%, 14.8% in 10% and 5% settings, respectively. Due to the scalability of Mask2Former and our simple yet effective training pipeline, which can fully utilize the open vocabulary knowledge from caption data, our model can perform well on open vocabulary panoptic segmentation.

Results on OVOD. Besides the segmentation task, we also evaluate Open Vocabulary Object Detection task by matching Ground Truth with bounding boxes in the testing stage. As shown in Tab. 3, CGG achieves better AP50 score in novel classes compared to several previous works [19, 79] in a shorter training schedule with no extra data used (only COCO-Caption). Previous methods like PromptDet [19] and Detic [79] tend to use large-scale image-text datasets, thus causing a longer training schedule and higher computational cost. We observe that CGG has inferior results in $AP50^{box}_{all}$. It may cause by the shorter training schedule and exposure to base classes compared with other methods.

4.3. Ablation Study and Analysis

We do ablation studies of our model to validate the effectiveness of each component. We do all the ablations on the MSCOCO 48/17 split [69] with the metric mAP.

Effectiveness of the CGG framework. First, we evaluate the significance of each proposed module. As shown in Tab. 4a, the baseline Class Emb., which transfers class labels to their corresponding text embeddings, only achieves a meager AP score of 0.2 for the novel class. As a comparison, adding Caption Grounding increases the Novel AP to 22.2, verifying the significance of Caption Grounding, which helps align multi-modal embeddings explicitly. Further, with the Caption Generation module, the final score reaches 28.4. The increase arises from a more strict regularization of Caption Generation, which supervises beyond nouns. For example, the caption "a woman **holding** an umbrella" requires the network to capture the relationship "**holding**" as well. If the Caption Generation module plays independently, the performance decreases to 0.3.

Training Pipeline. Previous methods like OVR-CNN [69] train their embeddings before fine-tuning the segmentor/detector, called 'emb-segm.' In this paper, we instead pre-train a class-agnostic segmentor and then train the multi-modal embeddings e_i^M on image-text data. We name it 'segm-emb.' These pipelines are compared over CGG in Tab. 4b. Note that "segm-emb-segm" is also included as a candidate. Results show that though "segm-emb" is inferior to others for base classes, it achieves significantly higher scores for novel classes. This phenomenon arises because training the segmentor in the last stage overfits the base classes, thus leading to a worse recall for novel classes.

Grounding Nouns Extraction. In CGG, we extract only object nouns from the sentences, leaving other words untouched. To validate the effectiveness of different word selection strategies, we also try to extract all words, and all nouns, except for only **object** nouns. The results are shown in Tab. 4c. By extracting all words, the novel AP decreases by 20.8, and by extracting all words without distinguishing whether they refer to objects or not, the novel AP decreases by 13.3. In conclusion, different extracting strategies significantly impact the model's performance.

Layers of Caption Generator. We evaluate the influence of layers of the transformer decoder in the caption generator. The results are in Tab. 4d. We test 2, 4, and 6 layers transformer decoders and observe that the middle number of 4 is a better choice, while fewer layers of 2 and more layers of 6 both harm the performance. Moreover, heavier caption generators may improve mAP for base classes a little, but also increase the computational cost.

Ablation on Class-Agnostic Pretraining. As a class-agnostic segmentor is trained to segment base and potential novel objects before training the multi-modal embeddings and caption generator, we do ablations on the effectiveness of class-agnostic pretraining and its alternatives. As shown in Tab. 4e, without any class-agnostic pretraining, the mAP on novel classes decreases by 5.7%. Moreover, if we freeze the Mask2Former, and only trains multi-modal embeddings and caption generator, the mAP on novel classes decreases by 2.0%, compared to the CGG model.

GFLOPs and Parameter Analysis. CGG adds a lightweight Transformer decoder as the caption generator in

Table 4. Ablation studies and comparison analysis on COCO OVIS.

(a) The Effectiveness of Each Components.

baseline	Gro.	Gen.	Base	Novel
Class Emb.			48.6	0.2
w. Gro.	✓		49.1	22.2
w. Gen.		✓	49.4	0.3
Both (CGG)	✓	✓	48.0	**28.4**

(b) Training Pipeline Comparison

Settings	Base	Novel	All
emb-segm	49.2	20.3	41.6
segm-emb-segm	**50.2**	24.3	**43.4**
segm-emb (CGG)	46.0	**28.4**	41.4

(c) Nouns Extraction in Caption Grounding

Method	Base	Novel	All
All Words	44.7	7.6	35.0
All Nouns	48.8	15.0	40.0
Object Nouns	46.0	**28.4**	41.4

(d) Caption Generator Design

#layers	Base	Novel	All
2	46.7	23.4	40.6
4 (CGG)	46.0	**28.4**	41.4
6	48.2	26.9	42.6

(e) Effect of Class-Agnostic Pretraining

Settings	Base	Novel	All
No class-agnostic	46.2	22.7	40.0
Freeze class-agnostic	47.6	26.4	42.1
CGG	46.0	**28.4**	41.4

(f) GFlops and Parameters

Schedule	Parameters	GFLOPs
baseline	35.65M	227.48
Ours: Inference	35.65M	227.48
Ours: Training	81.19M	229.33

Figure 5. Visualization of CGG results. Instance Segmentation (Top) and Panoptic Segmentation (Bottom). Categories marked by '*' are novel categories. Captions generated are depicted upon each image-prediction pair. Novel categories are colored in the caption if it has.

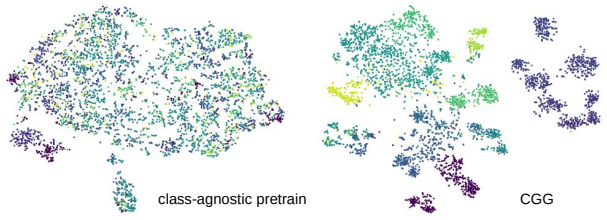

Figure 6. The embedding space of multi-modal embeddings $\{e_i^M\}$. The dimension of the embeddings is reduced to 2 dimensions using t-SNE [58]. Each color represents a class label in the 17 novel COCO classes. Each dot represents the embedding with the corresponding mask matched with ground truth annotations.

training. As shown in Tab. 4f, the #Parameters increases by 127.7% in training, while the total GFLOPs only increase by a small margin of 0.8%. Since text data is much smaller than images under the same batch size, the increased computational cost by the caption generator can be ignored. The GFLOPs and Parameters during inference are the same as the Mask2Former baseline.

Segmentation Results Visualization. Fig. 5 shows the qualitative results of CGG. The row shows panoptic results and the second row shows instance results. Novel classes detected in the image are marked with '*' and highlighted in the caption. The result shows that our framework can identify and segment base and novel classes. We show the generated comprehensive captions above the images.

Embeddings Space Visualization. Fig. 6 shows the t-SNE visualization result of the trained multi-modal embeddings (right) and the embeddings from a class-agnostic pretraining model (left). Specifically, we evaluate our model on the COCO validation set and get all the embeddings for each image. We change the bipartite matching strategy of Mask2Former to calculating mask loss only, thus getting the matching result between ground truth labels and multi-modal embeddings. The original Mask2Former [10] model cannot distinguish different novel classes in the embedding space, with only class-agnostic pretraining. After training using caption grounding and generation, the embeddings can formulate groups consistent with their categories.

5. Conclusion

In this paper, we present a joint Caption Grounding and Generation (CGG) framework for instance-level open vocabulary segmentation. Our core insights are two folds:

Firstly, the caption contains fine-grained nouns, which leads to better fine-grained grounding with object queries. Secondly, the caption can be a supervision signal that forces the model to predict novel objects. To our knowledge, we are the first to unify segmentation and caption generation for open vocabulary learning. We obtain significant performance improvement on both OVIS and OSPS and comparable results on OVOD *without* extra large-scale datasets pre-training.

Limitation and Future Work. Due to the limited computation resources, we do not pre-train our framework on extra caption datasets. Moreover, we do not use VLMs such as CLIP for distillation or supervision, and we do not experiment on larger scale datasets, like LVIS and Open-Image [25, 31]. We will put these as future work.

Acknowledgement. This work is supported by the National Key Research and Development Program of China (No.2020YFB2103402). We also gratefully acknowledge the support of SenseTime Research for providing the computing resources for this work.

6. More Implementation Details

Baseline Details. All the table results in main paper use **the same ResNet50 [27] backbone** for a fair comparison. The number of object queries is *100* by default. Our method is trained by only 12 epochs on the COCO train set and evaluated on the COCO validation set. All the experiments are carried out on 8 V100 GPUs. Following previous methods [28, 69], the metric we use for OVIS is mAP (mean AP on the IoU threshold of 0.5).

Training and Inference Details. We adopt the default training of Mask2Former [7, 10, 62]. 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 1024×1024. We use the default Mask R-CNN inference setting [26], where we resize an image with shorter side to 800 and longer side to 1333. *For the Inference of OSPS*, we do not use the default joint merge for things and stuff. We put the thing mask first and fill the remaining area with stuff mask prediction because the thing predictions for unknown are usually in a low score, and they may be covered by high score stuff mask prediction.

Training Splits For OVIS and OSPS. For OVIS, we follow the 48/17 split in COCO proposed by [48], in which 48 classes are base classes, and 17 are novel classes. For OSPS, we follow the unknown things split proposed by [29]. 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 5. 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

7. More Experiments Results

Will Joint Grounding and Caption Help the Fully Supervised Baseline? To answer this question, we perform ablation on fully supervised settings in Tab. 5. For the proposed CGG, we verify two main components, including caption grounding and caption 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. 5, 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. 5, for within class embedding settings, the added caption grounding and generation modules help to improve the performance on OVIS, but no performance gain on OSPS. We conclude that joint grounding and caption 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. Fig. 6 shows the results. As the caption generator becomes larger, the overall segmentation quality (AP all) increases. On the contrary, the quality of the caption (including BLUE and CIDEr) generation drops. This means a better caption generator may not be a better open vocabulary instance segmenter. The role of the cap-

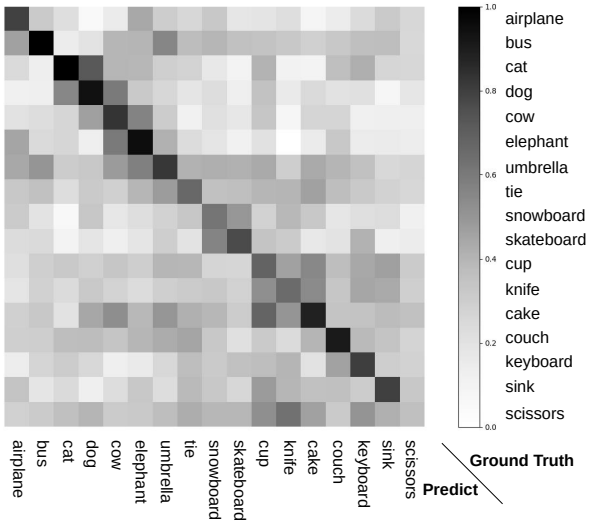

Figure 7. The correlation map between Ground Truth and model predictions on **novel classes**. The noun embeddings and object queries for novel classes are highly correlated.

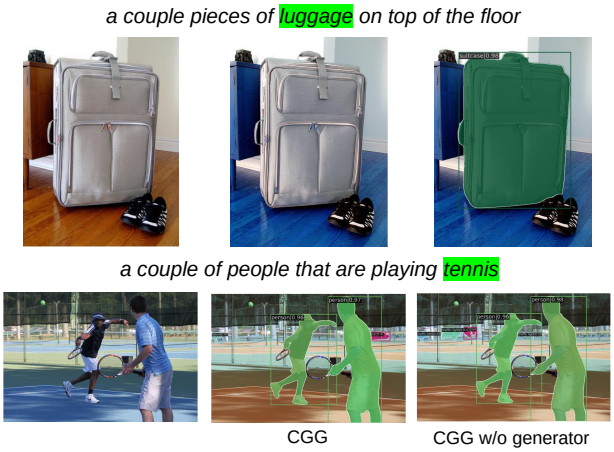

Figure 8. Examples of captions predicting objects that are not in the category list.

tion generator is to force the model to know the existence of novel objects, and pursuing a better caption generation model is not our goal of OVIS and OSPS.

8. Visual Analysis and Comparison

Visualization Analysis both Nouns and Object Queries. We calculate the correlation map between the predicted multi-modal embeddings e_i^M and the Ground Truth class embeddings. As shown in Fig. 7, 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 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 Fig. 8, images top. 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. 9 shows more visualization results with captions.

More Visualization Results on OVIS and OSPS. In Fig. 10, 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. 11, we show the visualization results on ADE20K dataset [77]. 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] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *ECCV*, pages 384–400, 2018. 2, 6

[2] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *ICCV*, 2019. 1

[3] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *NIPS*, 32, 2019. 1, 2

[4] Jiale Cao, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Sipmask: Spatial information preservation for fast image and video instance segmentation. *ECCV*, 2020. 1

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1, 4

[6] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. Blendmask top-down meets bottom-up for instance segmentation. In *CVPR*, 2020. 1

[7] 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. 6, 9

[8] Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollár. Tensormask: A foundation for dense object segmentation. In *ICCV*, 2019. 1

[9] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *CVPR*, 2020. 1

Table 6. 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

Figure 9. Visualization results of generated captions and the related segmentations of CGG. Input Image (Left), CGG (Middle), CGG w/o caption generation (Right). 'mirror' is not in the category list but is depicted by the generated caption.

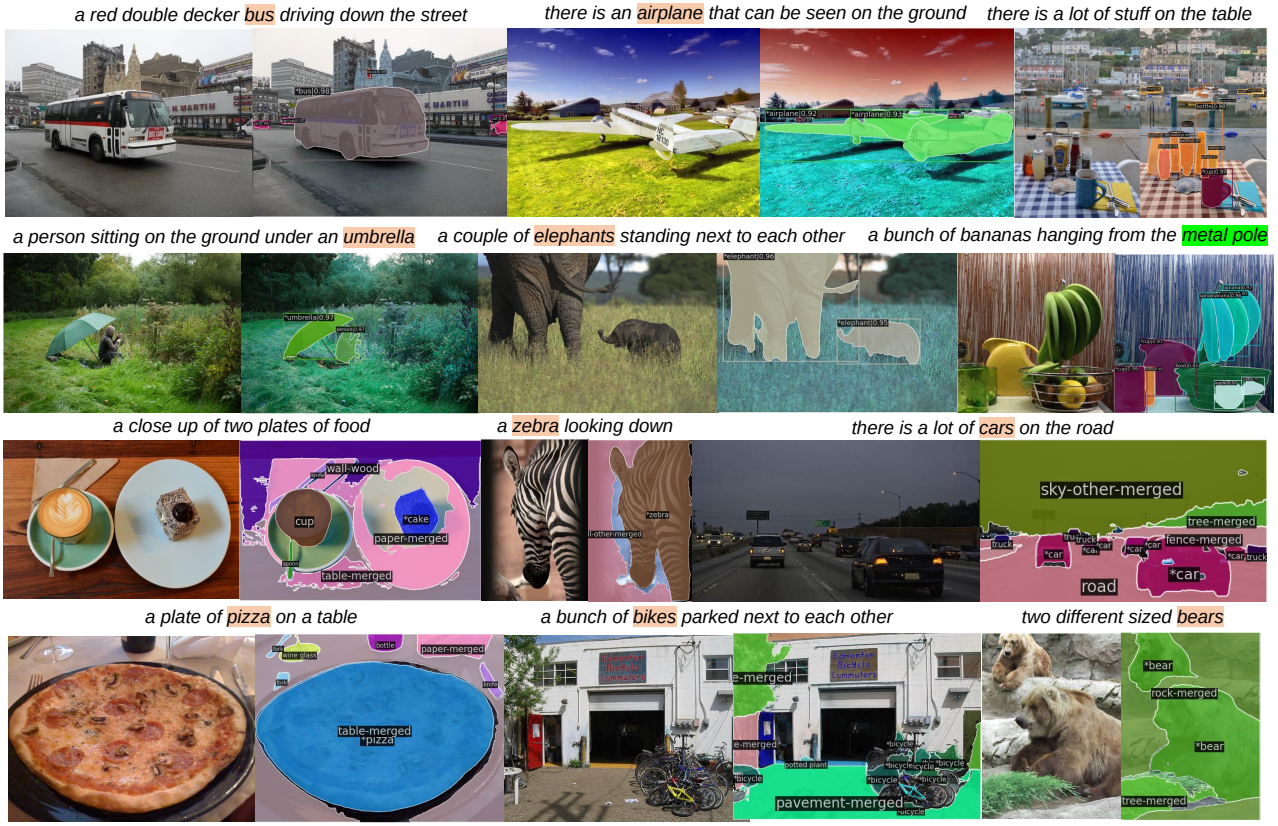

Figure 10. More visualization results of OVIS (Top two rows) and OSPS (Bottom two rows). Novel classes are marked by '*'.

[10] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. *CVPR*, 2022. 1, 2, 3, 4, 6, 8, 9, 12

Figure 11. Visualization on ADE20k [77]. Following [10], we apply instance segmentation on 100 instance classes. Classes not in COCO are marked by '*'.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 7

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[13] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, pages 16321–16330, 2021. 1

[14] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vlt: Vision-language transformer and query generation for referring segmentation. *TPAMI*, 2022. 1

[15] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *CVPR*, 2022. 1

[16] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, pages 14084–14093, 2022. 2, 3

[17] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 1

[18] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. *arXiv preprint arXiv:2105.01928*, 2021. 1

[19] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *ECCV*, 2022. 1, 6, 7

[20] A Frome, GS Corrado, J Shlens, et al. A deep visual-semantic embedding model. *NIPS*, pages 2121–2129, 2013. 2

[21] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, pages 540–557. Springer, 2022. 1, 2, 3, 4

[22] Cristina González, Nicolás Ayobi, Isabela Hernández, José Hernández, Jordi Pont-Tuset, and Pablo Arbeláez. Panoptic narrative grounding. In *ICCV*, pages 1364–1373, 2021. 1

[23] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 1, 2, 3

[24] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, pages 5356–5364, 2019. 1, 6

[25] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 9

[26] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 9

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 9

[28] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *CVPR*, pages 7020–7031, 2022. 1, 2, 3, 4, 6, 9

[29] Jaedong Hwang, Seoung Wug Oh, Joon-Young Lee, and Bohyung Han. Exemplar-based open-set panoptic segmentation network. In *CVPR*, pages 1175–1184, 2021. 2, 6, 7, 9

[30] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021. 1, 3

[31] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan

Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 1, 9

[32] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 3

[33] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 1

[34] Xiangtai Li, Shilin Xu, Yibo Yang, Guangliang Cheng, Yunhai Tong, and Dacheng Tao. Panoptic-partformer: Learning a unified model for panoptic part segmentation. In *ECCV*, 2022. 1

[35] Xiangtai Li, Ansheng You, Zhen Zhu, Houlong Zhao, Maoke Yang, Kuiyuan Yang, and Yunhai Tong. Semantic flow for fast and accurate scene parsing. In *ECCV*, 2020. 1

[36] Xiangtai Li, Wenwei Zhang, Jiangmiao Pang, Kai Chen, Guangliang Cheng, Yunhai Tong, and Chen Change Loy. Video k-net: A simple, strong, and unified baseline for video segmentation. In *CVPR*, 2022. 1

[37] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation. *CVPR*, 2021. 1

[38] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1

[39] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 1

[40] 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*, 2014. 1, 6

[41] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *ICCV*, pages 4673–4682, 2019. 1

[42] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017. 6

[43] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016. 3

[44] Davy Neven, Bert De Brabandere, Marc Proesmans, and Luc Van Gool. Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth. In *CVPR*, 2019. 1

[45] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019. 6

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1, 3

[47] Shafin Rahman, Salman Khan, and Nick Barnes. Improved visual-semantic alignment for zero-shot object detection. In *AAAI*, pages 11932–11939, 2020. 2

[48] Shafin Rahman, Salman Khan, and Fatih Porikli. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In *ACCV*, pages 547–563. Springer, 2018. 1, 2, 9

[49] Vignesh Ramanathan, Rui Wang, and Dhruv Mahajan. Dlwl: Improving detection for lowshot classes with weakly labelled data. In *CVPR*, pages 9342–9352, 2020. 6, 7

[50] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1

[51] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 1

[52] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018. 1

[53] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018. 6, 7

[54] Xing Shen, Jirui Yang, Chunbo Wei, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, Xiaoliang Cheng, and Kewei Liang. Dct-mask: Discrete cosine transform mask representation for instance segmentation. In *CVPR*, pages 8720–8729, 2021. 1

[55] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. SparseR-CNN: End-to-end object detection with learnable proposals. *CVPR*, 2021. 1

[56] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *CVPR*, 2020. 1

[57] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. *arXiv preprint arXiv:2003.05664*, 2020. 1

[58] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008. 8

[59] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015. 2, 3

[60] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. *arXiv preprint arXiv:2012.00759*, 2020. 1

[61] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. SOLOv2: Dynamic and fast instance segmentation. In *NeurIPS*, 2020. 1

[62] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 9

[63] Hai-Ming Xu, Hao Chen, Lingqiao Liu, and Yufei Yin. Two-stage decision improves open-set panoptic segmentation. *BMVC*, 2022. 2, 6

[64] Shilin Xu, Xiangtai Li, Jingbo Wang, Guangliang Cheng, Yunhai Tong, and Dacheng Tao. Fashionformer: A simple, effective and unified baseline for human fashion segmentation and recognition. *ECCV*, 2022. 1

[65] Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Cap2det: Learning to amplify weak caption supervision for object detection. In *ICCV*, pages 9686–9695, 2019. 7

[66] Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. Multimodal transformer with multi-view visual representation for image captioning. *TCSVT*, 30(12):4467–4480, 2019. 2, 3

[67] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, pages 1307–1315, 2018. 1

[68] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. *arXiv preprint arXiv:2203.11876*, 2022. 2

[69] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. *CVPR*, 2021. 1, 2, 3, 4, 5, 6, 7, 9

[70] Hui Zhang and Henghui Ding. Prototypical matching and open set rejection for zero-shot semantic segmentation. In *ICCV*, pages 6974–6983, 2021. 2

[71] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *arXiv preprint arXiv:2206.05836*, 2022. 1

[72] Rufeng Zhang, Zhi Tian, Chunhua Shen, Mingyu You, and Youliang Yan. Mask encoding for single shot instance segmentation. In *CVPR*, 2020. 1

[73] Wei Zhang, Wenbo Nie, Xinle Li, and Yao Yu. Image caption generation with adaptive transformer. In *YAC*, pages 521–526. IEEE, 2019. 2, 3

[74] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *CoRR*, abs/2106.14855, 2021. 1

[75] Wei Zhang, Yue Ying, Pan Lu, and Hongyuan Zha. Learning long-and short-term user literal-preference with multimodal hierarchical transformer network for personalized image caption. In *AAAI*, volume 34, pages 9571–9578, 2020. 3

[76] Ye Zheng, Jiahong Wu, Yongqiang Qin, Faen Zhang, and Li Cui. Zero-shot instance segmentation. In *CVPR*, pages 2593–2602, 2021. 6

[77] 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. 10, 12

[78] Chong Zhou, Chen Change Loy, and Bo Dai. Denseclip: Extract free dense labels from clip. *arXiv preprint arXiv:2112.01071*, 2021. 3

[79] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. *arXiv preprint arXiv:2201.02605*, 2022. 1, 2, 3, 6, 7

[80] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Don't even look once: Synthesizing features for zero-shot detection. In *CVPR*, pages 11693–11702, 2020. 2