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 open vocabulary instance segmentation to expand a segmentation model to classify and segment instance-level novel categories. Previous approaches have relied on massive caption datasets and complex pipelines to establish one-to-one mappings between image regions and words in captions. However, such methods build noisy supervision by matching non-visible words to image regions, such as adjectives and verbs. Meanwhile, context words are also important for inferring the existence of novel objects as they show high inter-correlations with novel categories. To overcome these limitations, we devise a joint ***Caption Grounding and Generation (CGG)*** *framework, which incorporates a novel grounding loss that only focuses on matching object nouns to improve learning efficiency. We also introduce a caption generation head that enables additional supervision and contextual modeling as a complementation to the grounding loss. Our analysis and results demonstrate that grounding and generation components complement each other, significantly enhancing the segmentation performance for novel classes. Experiments on the COCO dataset with two settings: Open Vocabulary Instance Segmentation (OVIS) and Open Set Panoptic Segmentation (OSPS) demonstrate the superiority of the CGG. Specifically, CGG achieves a substantial improvement of* ***6.8% mAP*** *for novel classes without extra data on the OVIS task and* ***15% PQ*** *improvements for novel classes on the OSPS benchmark.*

1. Introduction

Instance Segmentation [39] is a core vision task that goes beyond object detection [38, 37, 49] via segmenting and classifying each object. Despite it continues to attract significant research effort [25, 53, 12, 2, 56, 72, 13, 8, 4, 5, 7, 36, 35, 34, 66, 67], current solutions mainly focus on a closed-set problem that assumes a pre-defined set of object categories [39, 31, 23]. 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 [47, 3] is proposed, where models are trained on base classes and equipped with the ability to segment new classes. However, the zero-shot setting performs poorly on novel classes, as high-level word embeddings cannot effectively encode fine-grained visual information.

To address this issue, recent work [63] proposes an

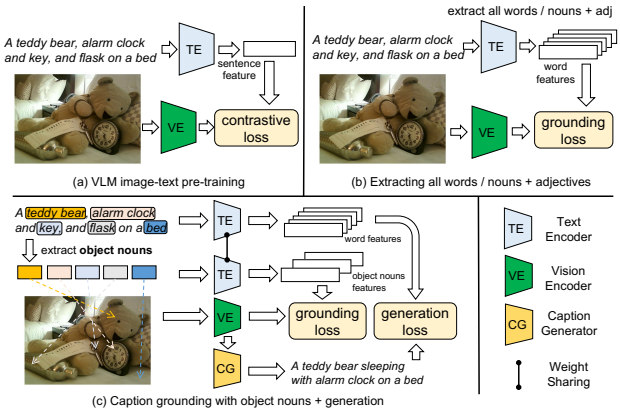

Figure 1: (a) VLMs learn image-level visual-linguistic alignment using caption data. (b) Previous open vocabulary detection/segmentation methods extract all words [63] or nouns + adjectives [20] for caption grounding. (c) The proposed CGG extracts object nouns for a finer alignment between objects in the caption and visible entities in the image and then combines a caption generation loss to utilize the contextual knowledge in the caption fully.

*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.

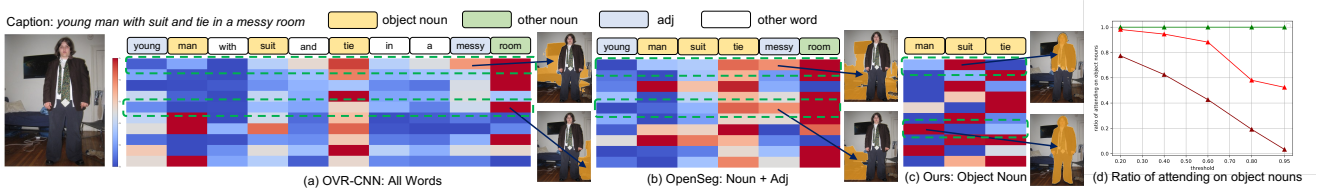

Figure 2: A comparison analysis of caption grounding using different types of words. The color maps are normalized similarities between multi-modal embeddings and word features extracted by the language encoder. Both (a) [63] and (b) [20] suffer from the problem that invisible nouns (room in the example) are learned to be aligned by the multi-modal embeddings while using object nouns avoids the question. We adopt top-10 object queries according to their object scores. (d) We sample 2500 images from the COCO validation set and test the average rate of multi-modal embeddings attending on object nouns under different thresholds.

open vocabulary setting by pre-training a visual backbone on captioned images for learning rich visual features. With the success of pre-trained Vision Language Models (VLMs) [45, 30], several approaches, *e.g.*, ViLD [22], 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 [20, 16]. Recently, state-of-the-art solutions [73, 18, 28, 33, 65] for open vocabulary detection/segmentation try to adopt larger-scale dataset pre-training with the help of VLMs. For example, Detic [73] adopts the ImageNet-21k [50] dataset to enlarge the detector in a weakly supervised manner, while PromptDet [18] augments the detection dataset with image-caption pairs scraped from the Internet. Recent XPM [28] also pre-trains their model on caption datasets [51]. These approaches typically require a complex architecture design to leverage extra datasets [50, 31]. 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 [61, 14, 15, 40, 12, 21] requires a model to align the text and corresponding region features, *e.g.*, OVR-CNN [63] and OpenSeg [20] in Fig. 1 (a) and (b). However, these methods expose a core issue in that they adopt the grounding loss between words and mask regions, implicitly assuming each word (or noun) should correspond to a region in the image. As shown in Fig. 2 (a) and (b), 'messy' and 'room' are forced to ground to meaningless masks. This motivates us to reformulate the ground loss by only focusing on object nouns as Fig. 2 (c). On the other hand, the latter [55, 60, 68] learns a model that outputs a caption for a given imagery input. It naturally captures the auxiliary and surrounding information to generate context words, which is crucial to building the bridge between image and text. Given the above observation, we argue that caption generation can naturally complement the grounding loss for context capturing.

Therefore, we propose a unified framework based on Mask2Former [8] performing each task jointly to exploit the knowledge from caption data better. It contains a caption grounding loss and an extra caption decoder for the generation loss, as shown in Fig. 1 (c). Motivated by the correlation analysis of object query and caption data (Sec. 3.2), we first extract object nouns for grounding loss. In particular, we transform the object queries into multi-modal embeddings using a linear layer at the input stage. Then we adopt separated object nouns to ground each multi-modal embedding, providing us with the grounding loss. Since extracted object nouns miss the structure information of caption data, we append a caption generation loss in the output stage to recover language data. We add a lightweight Transformer decoder with multi-modal embeddings as inputs to generate captions. Experiments demonstrate that the two losses are well coupled and mutually affect novel class segmentation, with only **0.8% GFlops** added during training. Our method drops the caption generation module for inference with no extra computation cost.

Our contributions can be summarized as follows:

- We propose a joint Caption Grounding and Generation (CGG) framework for open vocabulary instance segmentation, which incorporates grounding with object nouns and caption generating.
- Experimental results demonstrate our method achieves a significant improvement of **6.8% mAP** over previous XPM [28] on OVIS and **15% PQ** improvements over previous method [58] on OSPS.

2. Related Work

Zero-Shot Detection and Segmentation. Collecting and annotating data on a large scale is laborious and expensive for detecting and segmenting in an extensive vocabulary. Zero-Shot Detection [47] and Segmentation [3, 27, 26] aim to detect and segment novel categories that the annotations

are not accessible during training. To address this problem, many studies align region features with fixed text embeddings [19, 1, 46, 64, 74]. However, due to the limited capacity of word embeddings and the emergence of large Vision-Language-Models (VLMs), recent studies [63, 22, 62] have shifted towards the open vocabulary setting.

Open Vocabulary Object Detection (OVOD). Recent studies [17, 63, 22, 73, 62, 57] focus on the open vocabulary setting, where models are trained additionally on image-text pairs such as captions and text prompts. For example, OVR-CNN [63] pre-trains 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 pre-training on large-scale image-text pairs datasets. ViLD [22] distills the rich representation of pre-trained CLIP [45] into the detector, while DetPro [17] adds a fine-grained automatic prompt learning. Meanwhile, several works extract pseudo-region annotations from the pre-trained VLMs and use them as additional training data for detectors. Detic [73] improves the performance of the novel classes with image classification datasets by supervising the max-size proposal with various image labels. These methods above share a common idea of enlarging the capacity of training data to find rare classes, but they require more computation/annotation costs and complex pipelines. In contrast, we design a way to discover novel classes from caption data in one unified framework ***without*** pre-training on extra datasets nor distilling knowledge from pre-trained 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 into two different steps. The former generates mask regions, while the latter performs classification with pre-trained VLMs [20, 32]. DenseCLIP [71] proposes a similar pipeline to OVOD by distilling CLIP knowledge through generating pseudo mask labels. Our method proposes an end-to-end pipeline that jointly performs caption learning (grounding/generation) and segmentation learning. XPM [28] proposes a cross-modal pseudo-labeling framework by aligning word features in captions with visual features in images.

Image Captioning. This task requires the model to generate captions that describe the content of images [55]. State-of-the-art methods use multi-modal attention designs, treating the task as a multi-modal translation problem [60, 68, 69]. Our focus in this work is not on designing a new captioning model, but on exploring image captioning as a subtask for open vocabulary learning to enhance the novel class discovery ability. Using caption generation as an auxiliary loss is also adopted in vision language pre-training [9, 42]. However, to our knowledge, this is the *first study* exploring caption generation for OVS.

3. Methodology

In this section, we first review the background of OVIS and the baseline as preliminary. Then, we carry out the analysis on the correlation of caption data and query-based segmenter. Next, we present our Caption Grounding and Generation framework, which aims to exploit caption data via joint caption grounding and generation.

3.1. Preliminary

Problem Setting. We first describe the open-vocabulary problem setting. Let $\mathcal{D}_B = \{(I_m, M_m)\}_{m=1}^{N_B}$ 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. Each image I_m is associated with a set of ground-truth (GT) annotations M_m, which comprises instance masks and their corresponding object classes. To detect and segment novel classes, following previous works [63], we leverage additional image-level annotations, i.e., image captions. Let $\mathcal{D}_C = \{(I_c, C_c)\}_{c=1}^{N_C}$ be another set of training images with image caption annotations. Each image I_c is annotated with a caption C_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. OVIS aims to train a model to segment both base classes $\mathcal{V}_B$ and novel classes $\mathcal{V}_N$. Following previous methods [63, 28, 20], our model uses high-level semantic embeddings from a pre-trained text Transformer (BERT [11]) as the weights of the linear classifier. We focus on distilling knowledge in the captions to the target classes via representation similarities. In the following sections, we will neglect the image index for simplicity.

Baseline Method. We adopt the recent Mask2Former [8] model as our baseline since the query-based Transformer architecture can be readily extended into multi-modal training with captions. Given an image I, during the inference, Mask2Former directly outputs a set of M object queries $\boldsymbol{Q} = \{q_j | j = 1, .., M\}$, where each object query q_j represents one entity. Then, two different Multiple Layer Perceptrons (MLPs) project the queries into two embeddings for mask classification and prediction. During training, a bipartite matching algorithm matches each object query to the ground truth mask, following [8]. 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 [63], we use pre-trained embeddings to replace the learnable classifier for training and inference, as shown in Fig. 3. However, the original Mask2Former can only detect and segment closed-set objects and cannot handle the novel

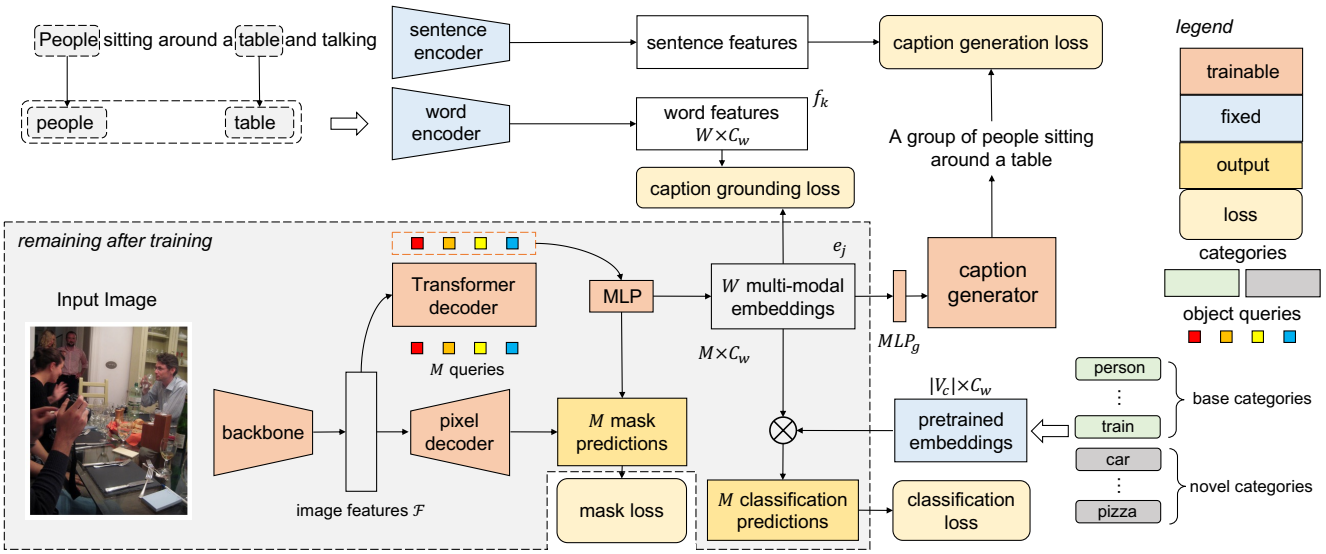

Figure 3: The illustration of **CGG** framework. The input image I is first provided to Mask2Former. The output of the Transformer decoder is then fed into an MLP, which generates M mask predictions together with the output of the pixel decoder in one hand. On the other hand, the object queries are transferred into M multi-modal embeddings, denoted as $\{e_j | j \in \{1, 2, \cdots, M\}\}$. The similarities of these embeddings with class embeddings are then computed to produce classification predictions. $\{e_j\}$ are also involved with grounding loss and generation loss with text features extracted by word and sentence encoder.

classes. Our method extends it to perform open-vocabulary segmentation in a new framework.

3.2. CGG Framework for OVS

Overview. Fig. 3 presents the overall pipeline of the CGG framework. Following [63], we set the pre-trained text embeddings as the weights of the linear classifier. Then 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, producing the image caption. During training, we adopt a pre-trained sentence encoder and word encoder to encode both captions and object nouns extracted from captions into sentence and word features. The former is used for caption generation, while the latter is for caption grounding. We discard all newly-introduced modules during inference and perform a lightweight inference procedure.

Analysis on Grounding Target with Object Query. Previous works like OVR-CNN [63] pre-train their models with caption data. However, 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 the segmenter is isolated, so 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.

For the first problem, we adopt a query-based detector [8] for end-to-end co-training. For the second problem, we argue that object nouns in caption data should be well aligned with query features in a more fine-grained manner since *the novel class categories are always nouns*. In Fig. 2 (a)-(c), we visualize the attention map of multi-modal embeddings and extracted word features, where we find several background items like rooms also have a high similarity with multi-modal embeddings, which brings the noise in supervision. In Fig. 2 (d), we perform a statistical analysis on the ratio of attended nouns, finding a significant drop with the increase of thresholds. Since object queries with higher scores always play as the output of instance segmentation, we argue this may hurt the performance of final segmentation results. Combining the above analysis and findings, we propose adopting object nouns as grounding targets.

Caption Grounding with Object Nouns. For the image-caption pair (I, C), we first extract object nouns from the caption C and feed it to the word encoder. Here, we neglect the image index for simplicity. We get word features $\{f_k | \ k \in \{1, 2, ..., K\}\}$, where K is the number of tokens from object nouns. For the image input I, we adopt an MLP layer to project the output of the Transformer decoder to a set of multi-modal embeddings $\{e_j | \ j \in \{1, 2, ..., M\}\}$, where M is the number of object queries in Mask2Former.

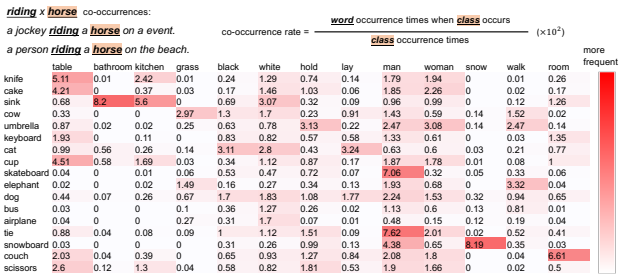

Figure 4: We often observe certain pairs of words co-occurrence while others do not. We calculate the co-occurrence matrix between novel classes and frequent words in the caption. Different classes have various distributions on co-occurrence words.

The similarity between image I and caption C is calculated as:

$$S^C(I,C) = \frac{1}{M}\sum_{j=1}^{M}\sum_{k=1}^{K} a^I_{j,k}\langle e_j, f_k\rangle, \tag{1}$$

where $\langle\cdot,\cdot\rangle$ is a dot production operation. $S^C(I,C)$ is normalized along the text dimension. $a^C_{j,k} = \exp\langle e_j, f_k\rangle / \sum_{l=1}^{K}\exp\langle e_j, f_l\rangle$ is the normalization term. Similarly, we can also get $S^I(I,C)$ by normalizing along the image dimension.

During training, the similarities between matching image-caption pairs should be maximized. For a mini-batch of image-caption pairs input $(\mathbf{I},\mathbf{C})$, the objective function is:

$$L^{CC}_{gro}(I) = -\log\frac{\exp S^C(I,C)}{\sum_{C'\in\mathbf{C}}\exp S^C(I,C')}, \tag{2}$$

and by normalizing along the image dimension, there is

$$L^{CI}_{gro}(I) = -\log\frac{\exp S^C(I,C)}{\sum_{I'\in\mathbf{I}}\exp S^C(I',C)}. \tag{3}$$

Similarly, we can get $L^{IC}_{gro}(I)$ and $L^{II}_{gro}(C)$ using $S^I(I,C)$. The final grounding loss for the batch is the summation of the four losses,

$$\begin{aligned} L_{gro} = &\frac{1}{|\mathbf{I}|}\sum_{I}(L^{CC}_{gro}(I) + L^{IC}_{gro}(I)) + \\ &\frac{1}{|\mathbf{C}|}\sum_{C}(L^{CI}_{gro}(C) + L^{II}_{gro}(C)). \end{aligned} \tag{4}$$

Grounding Object Nouns Misses the Structure Information of Caption Data. Despite grounding nouns forces to push the nouns embeddings and object queries closer, the structure information, including the relation of different objects is missing. As shown in Fig. 4, we perform co-occurrence relationship analysis on object nouns and find that novel classes have various distributions on co-occurrence words, which may help identify novel objects. This means only adopting grounding loss misses the relationship between these words. To fill such a gap, 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 grounding loss. As shown in Fig. 5, the caption generation module can help the model learn the specific status and relationships of objects in the scene.

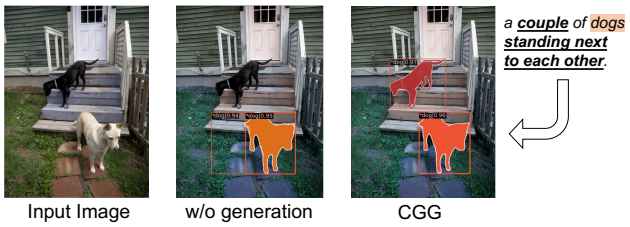

Figure 5: The effectiveness of caption generation. The generated caption depicts rich information beyond object nouns.

End-to-End Caption Generation Loss. Specifically, since the multi-modal embeddings encode the region-wise information, we transform these embeddings using a linear layer (MLP_g in Fig. 3) to fit the feature dimension of the caption generator, then we directly take the transformed embeddings as the input of the lightweight caption generator, which includes a stack of Transformer decoder layers. We adopt a Cross Entropy Loss on the predicted distribution of text vocabularies. It is the commonly used objective function in the research field of caption generation.

$$L_{gen} = -\sum_{t=1}^{N_s}\log(p_\theta(\hat{w}_t|w_1,...,w_{t-1})), \tag{5}$$

where $p_\theta(\hat{w}_t|w_1,\ldots,w_{t-1})$ is the predicted probability of t-th right word over the whole vocabulary, 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 $\{e_j|j \in \{1,2,\cdots,M\}\}$ capable of representing various words 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 [63], the classification loss is selected as the Cross-

Table 1: Results on Open Vocabulary Instance Segmentation.

Method	Constrained		Generalized		
	Base	Novel	Base	Novel	All
OVR [63]	42.0	20.9	41.6	17.1	35.2
SB [1]	41.6	20.8	41.0	16.0	34.5
BA-RPN [70]	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 COCO Open Vocabulary Object Detection (OVOD). IN-21K indicates ImageNet-21K [10]. CC indicates Conceptual Captions [52]

Method	Epochs	Extra Data	$AP50^{box}_{novel}$	$AP50^{box}_{all}$
DLWL [48]	96	YFCC100M	19.6	42.9
Cap2Det [59]	8.5	None	20.3	20.1
OVR-CNN [63]	12	None	22.8	39.9
Detic [73]	96	IN-21K & CC	24.1	44.7
PromptDet [18]	24	LAION-novel	26.6	50.6
CGG (Ours)	12	None	**29.3**	42.8

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.

Training and Inference. Compared to the baseline model, CGG only introduces extra losses and a caption generation head during the training. Following previous works [20, 63, 28], we first pre-train our framework using *only* base data annotations in a class-agnostic manner. The goal of pretraining is to encode instance-wised information into object queries. Then we load the pre-trained model for joint training with caption data. During the inference, following [63], we use the pre-trained embeddings of all classes to perform open vocabulary segmentation via dot product, including base classes and novel classes.

4. Experiments

4.1. Experimental Setup

Dataset Settings. We conduct experiments on COCO dataset [39] for OVIS and OSPS. For OVIS, following previous works [63, 28], we split 48 base classes with annotations and 17 target classes without annotations. For captioned images, we use the entire COCO-captions training set with 118,287 images and five captions per image. Unlike previous works [73, 18, 48] that adopt extra caption datasets, like Conceptual Captions [52] for pre-training, we do **not** use extra caption or detection datasets. We follow the origin OVR-CNN [63] setting by only exploring a limited caption dataset within COCO. For OSPS [29], we follow the previous works [29, 58], 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%).

Table 3: We compare our method CGG with previous methods EOPSN [29] and Dual [58] on Open Set Panoptic Segmentation (OSPS). Unlike EOPSN and Dual that group all unknown things into one class without identifying them, CGG performs Open Vocabulary Panoptic Segmentation and assigns a specific category to each unknown thing. We show the mean PQ and SQ for all unknown categories and indicate the scores averaged from each unknown class with "*".

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 [58]		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 [29]	10	44.5	80.6	28.4	71.8	17.9	76.8
Dual [58]		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 [29]	20	45.0	80.3	28.2	71.2	11.3	73.8
Dual [58]		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*

Metric. For OVIS, we report the mask-based mean Average Precision (mAP) at intersection-over-union (IoU) of 0.5. Following previous works [63, 28], we evaluate the model performance on base and target classes in two settings: constrained setting, where the model is only tested on images that belong to either base or target classes; generalized setting, where the model is tested on both base and target classes. The latter is more challenging, as it requires the model to avoid class bias from base classes. We also report open vocabulary detection with box-based mAP. For OSPS setting, we use panoptic segmentation metrics, including Panoptic Quality (PQ) and Segmentation Quality (SQ). 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 [44] with MMDetection framework [6]. We use 8 GPUs for distributed training. Each mini-batch has two images per GPU. The optimizer is AdamW [41] with a weight decay of 0.0001. We adopt full image size for a random crop in the pre-training and training process following [8]. We use BERT embeddings [11] for the classification head, word encoder, and sentence encoder. We use an LVIS class name parser to extract object nouns from caption data. For OVIS, we keep the top 100 queries as the model outputs. For OSPS, we follow previous works [29, 58], which put thing mask predictions first, then fill the remaining back-

Table 4: Ablation studies and 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
Nouns + Adj	46.2	16.2	39.2
Object Nouns + Adj	45.6	27.2	40.2
Object Nouns	46.0	**28.4**	41.4

(d) Caption Generator Design

#layers	Base	Novel	All
2	46.7	23.4	40.6
4	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

Table 5: Ablation on the ability of caption generation.

#layers	BLUE-1 ↑	BLUE-2 ↑	BLUE-3 ↑	BLUE-4 ↑	CIDEr ↑	ROUGE ↑
2	0.473	0.311	0.206	0.141	0.307	0.360
4	0.418	0.258	0.166	0.111	0.239	0.320
6	0.387	0.226	0.138	0.088	0.171	0.289

ground with stuff mask predictions. We use the ResNet-50 backbone for all experiments for a fair comparison.

4.2. Main Results

Results on OVIS. We first compare CGG and other methods for the OVIS task. Tab. 1 shows that our model outperforms XPM, the best baseline, by 5.5% mAP in the constrained setting 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 must distinguish novel categories from base categories, where the training data bias is for base categories. CGG has improved more in generalized than constrained settings, demonstrating its effectiveness in identifying and distinguishing novel classes from base classes.

Results on OVOD. We further evaluate our model on the Open Vocabulary Object Detection task, which requires matching ground truth with predicted bounding boxes at test time. Tab. 2 shows that CGG outperforms several previous works [18, 73] on novel classes in terms of AP50 score while using only COCO-Captions as the image-text data source and a shorter training schedule. Previous methods such as PromptDet [18] and Detic [73] rely on large-scale image-text datasets, which incur a longer training time and higher computational cost. However, CGG performs worse on all classes cases: $AP50^{box}_{all}$. It may be due to the limited exposure to base classes and shorter training schedules compared with other methods.

Results on OSPS. We test CGG on the Open Set Panoptic Segmentation task by expanding the base classes from base thing classes to including stuff classes without changing the training pipeline of CGG. Tab. 3 indicates that our model performs better than previous methods EOPSN [29] and Dual [58] by 14.9% PQ on unknown things in 20% unknown things setting, and 16.9%, 14.8% in 10% and 5% settings, respectively. Compared with the standard Open Set Panoptic Segmentation task, CGG classifies each unknown class and still outperforms previous methods.

Table 6: Comparison between only training caption generation and joint training with segmentation. "only-gen" means the model is trained purely with caption generation supervision.

Method	BLUE-1 ↑	BLUE-2 ↑	BLUE-3 ↑	BLUE-4 ↑	CIDEr ↑	ROUGE ↑
only gen.	0.394	0.237	0.150	0.100	0.177	0.305
CGG	**0.418**	**0.258**	**0.166**	**0.111**	**0.239**	**0.320**

4.3. Ablation Study and Analysis

To evaluate the effectiveness of each component of our model, we conduct ablation studies on the COCO 48/17 split [63] using mAP as the metric.

Effectiveness of Modules. We first verify the effectiveness of each proposed module in CGG. Tab. 4a shows that the baseline Class Emb., which maps class labels to text embeddings, achieves a low AP score of 0.2 for the novel class. By contrast, adding Caption Grounding boosts the Novel AP to 22.2, demonstrating the importance of Caption Grounding for aligning multi-modal embeddings to object nouns. The final score reaches 28.4. This improvement comes from Caption Generation, which supervises object nouns and other meaningful words. Without Caption Grounding, Caption Generation alone performs poorly with 0.3 AP. This observation demonstrates that Caption Grounding is the crucial module.

Training Pipeline. We compare different training pipelines

Figure 6: Visualization results on Instance Segmentation (Top) and Panoptic Segmentation (Bottom). The categories with "*" are novel. We also generate captions for each image-prediction pair and highlight the novel categories in the captions, if any.

for our model. Previous methods like OVR-CNN [63] use an "emb-segm" pipeline, which trains with captions first and then fine-tunes the segmentor. In contrast, we adopt a "segm-emb" pipeline, which pre-trains a class-agnostic segmentor and then trains the multi-modal embeddings e_i^M on image-text data. Tab. 4b compares these pipelines over CGG. We also include "segm-emb-segm" as a candidate. The results indicate that although "segm-emb" performs worse than others for base classes, it achieves much higher scores for novel classes. The lower performance for base classes is because CGG is first pre-trained on only base classes and then fine-tuned jointly with captions. Thus, the performance increases for novel classes while drops for base classes since the segmenter overfits the base classes during the first stage. A possible solution is to balance the ratio of base and novel classes during fine-tuning. Training the segmentor in the last stage causes overfitting on the base classes and reduces recall for novel classes.

Grounding Nouns Extraction. We investigate different word selection strategies for CGG as discussed in Sec. 3.2. Instead of extracting only object nouns from the sentences, we extract all words [63], nouns + adj[20], and object nouns + adj. Tab. 4c shows the results of these strategies. Extracting all words leads to a 20.8 drop in novel class AP, and extracting nouns + adj leads to a 12.2 drop in novel class AP. When extracting object nouns + adj, the performance is close to ours. These results indicate that selecting suitable words is crucial for our model's performance, where we find object nouns perform the best.

Layers of Caption Generator. We examine the effect of numbers in the Transformer decoder layers in the caption generator. Tab. 4d shows the results for 2, 4, and 6 layers. Adding more layers cannot always boost the performance of novel class AP. However, the AP score for all classes increases when the caption generator becomes larger. It is because base categories occur more frequently than novel categories, and benefit more from model enlargement.

Ability of Caption Generation. We also explore the generated caption quality, despite this is not our goal for OVIS. In Tab. 5, we observe that adding more Transformer layers in the caption generator cannot improve the model's ability of caption generation. However, in Tab. 6, we train the model with only caption generation supervision and get lower generation scores than the joint training. These results indicate that multitask training may also improve the effectiveness of the generation task.

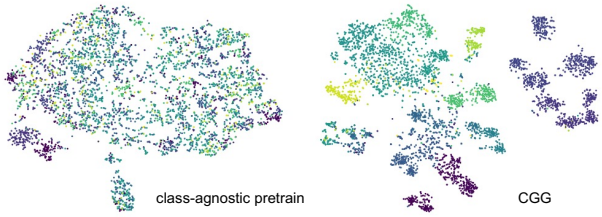

Figure 7: The multi-modal embeddings e_i^M in a 2D space using t-SNE [54]. The colors indicate the class labels of the 17 novel COCO classes. The dots represent the embeddings with the masks that match the ground truth annotations.

Ablation on Class-Agnostic Pretraining. We investigate the effect of class-agnostic pre-training on our model. The class-agnostic model is trained to segment base and potential novel objects before training the multi-modal embeddings and caption generator. Tab. 4e reports the results of different pre-training strategies. Without class-agnostic pre-training, the mAP on novel classes drops by 5.7%. If fixing Mask2Former in pre-training and only training multi-modal embeddings and caption generator, the mAP on novel classes drops by 2.0%. This indicates end-to-end training plays an important role for query-based segmenter.

GFLOPs and Parameter Analysis. CGG introduces a lightweight Transformer decoder as the caption generator during training. As shown in Tab. 4f, this increases the number of parameters by 127.7% in training, while the total GFLOPs increase only by 0.8%. Since text data is much

smaller than images under the same batch size, the additional computational cost brought by the caption generator can be ignored. The GFLOPs and Parameters during inference are the same as the Mask2Former baseline.

Segmentation Results Visualization. We present some qualitative results of CGG in Fig. 6. The first row shows panoptic results, and the second shows instance results. Novel classes are marked with "*" and highlighted in the caption. The result demonstrates that our framework can segment and identify base and novel classes. We also show the generated comprehensive captions above.

Embeddings Space Visualization. We visualize the multi-modal embeddings learned by CGG and a class-agnostic pretraining baseline using t-SNE in Fig. 7. We extract the predicted embeddings for each image in the COCO validation set and match them with ground truth labels by mask similarity. The baseline model fails to cluster the embeddings by their categories due to the lack of class-specific knowledge. In contrast, CGG leverages caption grounding and generation to learn discriminative embeddings that align with their semantic classes.

5. Conclusion

This paper presents a joint Caption Grounding and Generation (CGG) framework for instance-level open vocabulary segmentation. The main contributions are: (1) using fine-grained object nouns in captions to improve grounding with object queries. (2) using captions as supervision signals to extract rich information from other words helps identify novel categories. To our knowledge, this paper is the first to unify segmentation and caption generation for open vocabulary learning. The proposed framework significantly improves OVIS and OSPS and comparable results on OVOD *without* pre-training on large-scale datasets.

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 OpenImage [24, 31]. We will put these as future work.

Acknowledgement. This work is supported by the National Key Research and Development Program of China (No.2020YFB2103400). This study also is supported under the RIE2020 Industry Alignment Fund Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). It is also supported by Singapore MOE AcRF Tier 2 (MOE-T2EP20120-0001). We also gratefully acknowledge the support of SenseTime Research for providing the computing resources for this work.

References

[1] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *ECCV*, 2018. 3, 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. *NeurIPS*, 2019. 1, 2

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

[5] 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

[6] 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

[7] 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

[8] 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

[9] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, 2021. 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*. IEEE, 2009. 6

[11] 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, 6

[12] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, 2023. 1, 2

[13] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. MOSE: A new dataset for video object segmentation in complex scenes. In *ICCV*, 2023. 1

[14] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, 2021. 2

[15] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. VLT: Vision-language transformer and query generation for referring segmentation. *TPAMI*, 2023. 2

[16] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *CVPR*, 2022. 2

[17] 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*, 2022. 3

[18] 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. 2, 6, 7

[19] A Frome, GS Corrado, J Shlens, et al. A deep visual-semantic embedding model. *NeurIPS*, 2013. 3

[20] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*. Springer, 2022. 1, 2, 3, 6, 8

[21] Cristina González, Nicolás Ayobi, Isabela Hernández, José Hernández, Jordi Pont-Tuset, and Pablo Arbeláez. Panoptic narrative grounding. In *ICCV*, 2021. 2

[22] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *ICLR*, 2022. 2, 3

[23] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 1

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

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

[26] Shuting He, Henghui Ding, and Wei Jiang. Primitive generation and semantic-related alignment for universal zero-shot segmentation. In *CVPR*, 2023. 2

[27] Shuting He, Henghui Ding, and Wei Jiang. Semantic-promoted debiasing and background disambiguation for zero-shot instance segmentation. In *CVPR*, 2023. 2

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

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

[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*. PMLR, 2021. 2

[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, 2, 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. 2

[34] Xiangtai Li, Henghui Ding, Wenwei Zhang, Haobo Yuan, Guangliang Cheng, Pang Jiangmiao, Kai Chen, Ziwei Liu, and Chen Change Loy. Transformer-based visual segmentation: A survey. *arXiv pre-print*, 2023. 1

[35] Xiangtai Li, Haobo Yuan, Wenwei Zhang, Guangliang Cheng, Jiangmiao Pang, and Chen Change Loy. Tube-link: A flexible cross tube baseline for universal video segmentation. *ICCV*, 2023. 1

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

[37] 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

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

[39] 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

[40] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *ICCV*, 2019. 2

[41] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[42] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 3

[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] 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

[45] 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*, 2021. 2, 3

[46] Shafin Rahman, Salman Khan, and Nick Barnes. Improved visual-semantic alignment for zero-shot object detection. In *AAAI*, 2020. 3

[47] Shafin Rahman, Salman Khan, and Fatih Porikli. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In *ACCV*, 2018. 1, 2

[48] Vignesh Ramanathan, Rui Wang, and Dhruv Mahajan. Dlwl: Improving detection for lowshot classes with weakly labelled data. In *CVPR*, 2020. 6

[49] 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

[50] 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. 2

[51] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 2

[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*, 2018. 6

[53] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. *ECCV*, 2020. 1

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

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

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

[57] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, Bernard Ghanem, and Dacheng Tao. Towards open vocabulary learning: A survey. *arXiv pre-print*, 2023. 3

[58] Hai-Ming Xu, Hao Chen, Lingqiao Liu, and Yufei Yin. Two-stage decision improves open-set panoptic segmentation. *BMVC*, 2022. 2, 6, 7

[59] 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*, 2019. 6

[60] Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. Multimodal transformer with multi-view visual representation for image captioning. *TCSVT*, 2019. 2, 3

[61] 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*, 2018. 2

[62] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. *arXiv preprint arXiv:2203.11876*, 2022. 3

[63] 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, 8

[64] Hui Zhang and Henghui Ding. Prototypical matching and open set rejection for zero-shot semantic segmentation. In *ICCV*, 2021. 3

[65] 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. 2

[66] Jiangning Zhang, Xiangtai Li, Jian Li, Liang Liu, Zhucun Xue, Boshen Zhang, Zhengkai Jiang, Tianxin Huang, Yabiao Wang, and Chengjie Wang. Rethinking mobile block for efficient neural models. *ICCV*, 2023. 1

[67] Jiangning Zhang, Chao Xu, Jian Li, Wenzhou Chen, Yabiao Wang, Ying Tai, Shuo Chen, Chengjie Wang, Feiyue Huang, and Yong Liu. Analogous to evolutionary algorithm: Designing a unified sequence model. *NeurIPS*, 2021. 1

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

[69] 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*, 2020. 3

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

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

[72] Qianyu Zhou, Xiangtai Li, Lu He, Yibo Yang, Guangliang Cheng, Yunhai Tong, Lizhuang Ma, and Dacheng Tao. Transvod: End-to-end video object detection with spatial-temporal transformers. *T-PAMI*, 2023. 1

[73] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. *ECCV*, 2022. 2, 3, 6, 7

[74] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Don't even look once: Synthesizing features for zero-shot detection. In *CVPR*, 2020. 3