Global Knowledge Calibration for Fast Open-Vocabulary Segmentation

Kunyang Han[1,2] *, Yong Liu[3] *, Jun Hao Liew[4], Henghui Ding[5], Jiajun Liu[4]
Yitong Wang[4], Yansong Tang[3], Yujiu Yang[3], Jiashi Feng[4], Yao Zhao[1,2] ✉, Yunchao Wei[1,2]

[1] Institute of Information Science, Beijing Jiaotong University
[2] Beijing Key Laboratory of Advanced Information Science and Network Technology
[3] Tsinghua Shenzhen International Graduate School, Tsinghua University
[4] ByteDance   [5] Nanyang Technological University

Abstract

Recent advancements in pre-trained vision-language models, such as CLIP, have enabled the segmentation of arbitrary concepts solely from textual inputs, a process commonly referred to as open-vocabulary semantic segmentation (OVS). However, existing OVS techniques confront a fundamental challenge: the trained classifier tends to overfit on the base classes observed during training, resulting in suboptimal generalization performance to unseen classes. To mitigate this issue, recent studies have proposed the use of an additional frozen pre-trained CLIP for classification. Nonetheless, this approach incurs heavy computational overheads as the CLIP vision encoder must be repeatedly forward-passed for each mask, rendering it impractical for real-world applications. To address this challenge, our objective is to develop a fast OVS model that can perform comparably or better without the extra computational burden of the CLIP image encoder during inference. To this end, we propose a core idea of preserving the generalizable representation when fine-tuning on known classes. Specifically, we introduce a text diversification strategy that generates a set of synonyms for each training category, which prevents the learned representation from collapsing onto specific known category names. Additionally, we employ a text-guided knowledge distillation method to preserve the generalizable knowledge of CLIP. Extensive experiments demonstrate that our proposed model achieves robust generalization performance across various datasets. Furthermore, we perform a preliminary exploration of open-vocabulary video segmentation and present a benchmark that can facilitate future open-vocabulary research in the video domain.
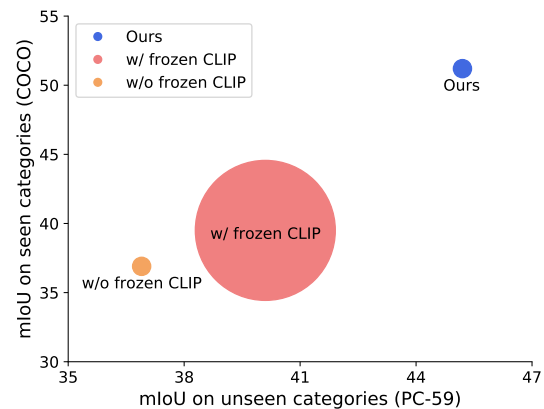
Figure 1. **Performance *vs*. computational cost**. The radius of the circle represents the FLOPs during inference. To avoid overfitting to the seen categories, some methods [14, 44] introduce an **extra** frozen CLIP during inference. However, such a strategy leads to heavy computation overhead (red ●). In comparison, our method generalizes well on both seen and unseen categories with much smaller computational cost (blue ●).

1. Introduction

Semantic segmentation aims to group pixels that belong to the same categories. Despite achieving high performance in recent years [30, 11, 38, 42, 4, 19, 16, 36, 48, 49], existing semantic segmentation approaches often rely on predefined sets of training categories and thus cannot recognize categories that were not present during training. This limitation greatly restricts their practical applicability. In contrast, humans possess the ability to recognize novel categories in an open-vocabulary manner, *i.e.*, identifying objects using arbitrary text from an unbounded vocabulary. This ability has inspired the development of open-vocabulary segmentation methods [14, 44, 47, 21, 18, 41, 25]. Unlike traditional closed-set segmentation, open-vocabulary segmentation can segment arbitrary categories given only text inputs,

---

*Equal contribution. Work done during internships at ByteDance.
✉Corresponding author.

which has many potential applications, such as image editing and human-robot interaction.

To achieve open-vocabulary segmentation, early approaches [41, 47, 25] replace the output classification layer with cross-modal alignment, where the similarity measure between pixels and text embeddings is used. Recent works [14, 18, 28, 22, 44], on the other hand, adopt the region-level alignment approach and have demonstrated remarkable performance. Despite these advancements, open-vocabulary segmentation methods still face a significant challenge: the learned embeddings often overfit to the base classes observed during training, which hinders their ability to generalize to novel classes. To overcome this challenge, some methods [44, 14, 28] utilize an additional frozen CLIP vision encoder for re-classification. However, this strategy incurs heavy computation overhead, as it requires repeated forward passes of the CLIP vision encoder for each mask. This can be prohibitively expensive for real-world applications, as illustrated in Fig. 1.

Therefore, our objective is to train an open-vocabulary semantic segmentation model that is fast and does not require the extra heavy CLIP image encoder during inference, while achieving comparable or better performance. The two main factors that contribute to this objective are: (1) the model should not overfit to the specific training category names, and (2) the model should maintain a feature space similar to the pre-trained CLIP. To achieve this goal, we introduce **Global Knowledge Calibration**. To prevent the learned representation from being biased towards the specific training category names, we propose a text diversification strategy for prompt augmentation. This strategy enhances text diversity and enriches category semantics with information of different granularities. Specifically, we use WordNet [17] to generate a set of synonyms for each training category, *e.g.*, "vessel" and "ship" for "boat", and expand the initial text prompts with this set of words.

To maintain the generalizable knowledge of CLIP [37], a straightforward solution is to apply knowledge distillation. However, traditional knowledge distillation methods only utilize the CLIP features of the same object as supervision. As a result, they can only fit the representations of individual classes and fail to effectively model the overall CLIP space. To address this issue, we propose a text-guided knowledge distillation strategy for calibrating the representation of the trained model. Specifically, we apply distillation supervision for the visual embeddings of one category by using all categories present in the image. Using the distance between category names in the text space as guidance, this distillation strategy can guide the trained model to build a multi-modal feature space similar to the pre-trained CLIP.

In addition, to our best knowledge, previous research on OVS has only focused on the image domain. In this work, we make a preliminary exploration of open-vocabulary video segmentation. We introduce a benchmark by partitioning the large-scale video segmentation dataset, VIPSeg [32], into seen and unseen categories for zero-shot testing. We develop a simple baseline based on our image-based method. Our aim is to provide support for future open-vocabulary research in the video domain.

Our contributions can be summarized as follows:

- We propose Global Knowledge Calibration to preserve generalizable representations when training solely on known classes. Our approach does not require an additional heavy CLIP vision encoder during inference, making it faster. Extensive experiments demonstrate that our model offers strong generalization performance across various datasets, with a much smaller computational cost.

- We present a text diversification strategy to enrich text supervision with category information of varying granularities. We propose a text-guided knowledge distillation strategy to calibrate the learned feature space.

- To the best of our knowledge, we are the first to explore open-vocabulary video segmentation. We construct a new benchmark and a simple baseline.

2. Related Work

Vision-Language Pre-training. Vision-language pre-training aims to learn a joint visual-textual representation space. Early approaches [5, 26, 27, 31, 35] were limited to small-scale datasets and required fine-tuning on downstream tasks. With the availability of large-scale web data, recent works [24, 37] have demonstrated the benefits of utilizing such data to learn a more robust multi-modal representation space. CLIP [37] leverages the idea of contrastive learning to connect images with their corresponding captions and has achieved impressive cross-modal alignment performance. Inspired by previous works [14, 25, 44], we utilize the well-aligned space of CLIP to enhance open-vocabulary segmentation tasks.

Open-Vocabulary Segmentation. The open-vocabulary segmentation task aims to segment an image and identify regions with arbitrary text queries [18, 1]. Pioneering work by ZS3Net [1] proposed training a generator to synthesize visual representations by transforming word embeddings. With the generator expanding the pseudo unseen class visual features, the classifier is trained to distinguish between real features from seen categories and synthetic features from unseen categories. SPNet [41] replaces the prediction convolution layer by computing the similarity between visual features and linguistic embeddings, while GroupViT [43] learns to group image regions by contrastive learning between text and images. LSeg [25] proposes
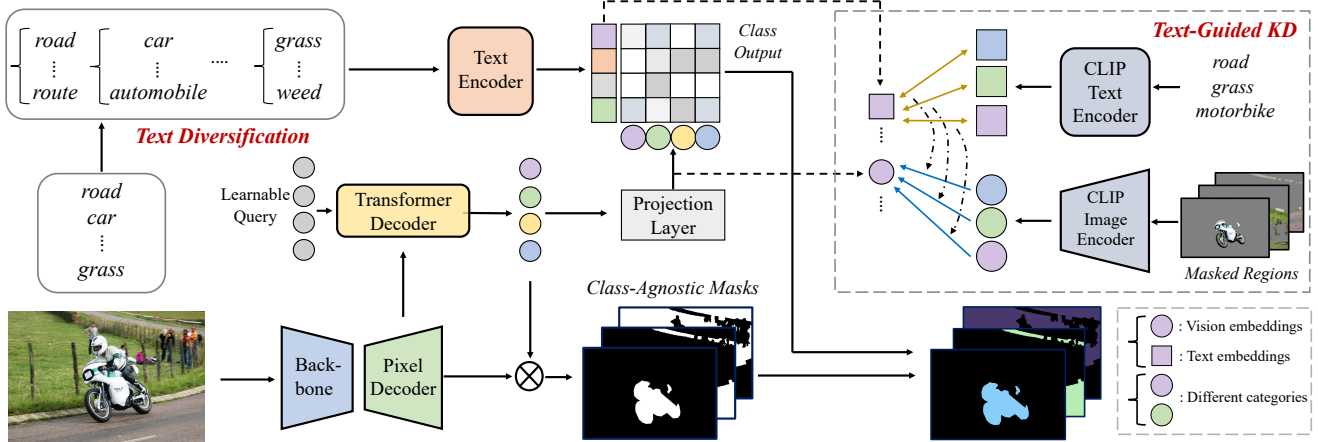
Figure 2. **Pipeline**. The input image is first encoded into hierarchical features by visual backbone and pixel decoder. A transformer decoder takes the hierarchical features and a group of learnable queries as input and outputs visual region queries. By perceiving image content, the region queries contain the information of different category regions in the image. By combining the region queries with hierarchical visual features, the model can generate class-agnostic mask proposals. Simultaneously, the region queries are projected towards the textual space with a projection layer. By calculating the similarity between the region queries with the text embeddings of each category name, the model outputs the classification prediction for each mask. During training, we apply both the **text diversification strategy** and **text-guided knowledge distillation** to improve the representation of visual and textual embeddings.

maximizing the correlation between the language embedding and visual pixel-level embeddings using a pre-trained CLIP [37] text encoder. More recently, a two-stage pipeline was proposed: the model first generates class-agnostic region proposals, followed by segment-level alignment between proposals and linguistic embeddings. OpenSeg [18] leverages a segmentation model to divide input images into regions and computes the grounding loss between the regions and text. Simbaseline [44] crops the input image based on the proposal masks and utilizes CLIP [37] to extract region-level features. Afterward, the segment embeddings are classified by computing similarity with category name embeddings. Zegformer [14] uses CLIP as the encoder and MaskFormer [8] to extract mask proposals. However, both Simbaseline and Zegformer require an extra CLIP image encoder to extract the instance embeddings according to the proposal masks, increasing the model parameters and complicating the inference process. To address these issues and further improve the performance of open-vocabulary segmentation, we propose Global Knowledge Calibration in this paper.

3. Global Knowledge Calibration

Pipeline. As depicted in Fig. 2, our method utilizes a "segment-then-classify" pipeline for open-vocabulary segmentation task. Initially, the input image is encoded into hierarchical visual features by a visual backbone and a pixel decoder. Subsequently, a transformer [39, 2] decoder takes a set of learnable queries and hierarchical visual features as input to generate region-aware queries (indicated

by colored circles in the figure). Next, the region-aware queries are fused with the output of the pixel decoder to produce class-agnostic masks. Concurrently, the region-aware visual queries are fed into a projection layer to perform cross-modal alignment with textual embeddings. The alignment score represents the classification confidence of each query. By combining the class-agnostic masks with the cross-modal alignment scores, our model assigns categories to each mask based on the maximum score. For cross-modal alignment, we use textual embeddings generated by a frozen text encoder [37] that takes category names with prompt templates as input. Notably, unlike conventional approaches that rely on the initial class name defined in training datasets, we propose a text diversification strategy to enhance text diversity (Sec. 3.1). Specifically, we leverage WordNet [17] to generate a set of synonyms for each category name, and perform cross-modal matching on all synonyms with corresponding scores. Furthermore, given the high generalizability of pre-trained CLIP [37] space, we propose a text-guided knowledge distillation strategy to maintain the CLIP representation even for unseen categories (Sec. 3.2).

3.1. Text Diversification Strategy

Using only category names as text prompts during training can result in overfitting to specific words and limit the model's ability to generalize. To overcome this limitation, we propose a text diversification strategy that enriches the text prompts with different words that have similar meanings. To achieve this, we leverage WordNet [17] to generate a set of synonyms $w_i^0, w_i^1, \ldots, w_i^{N_i}$ for each category name
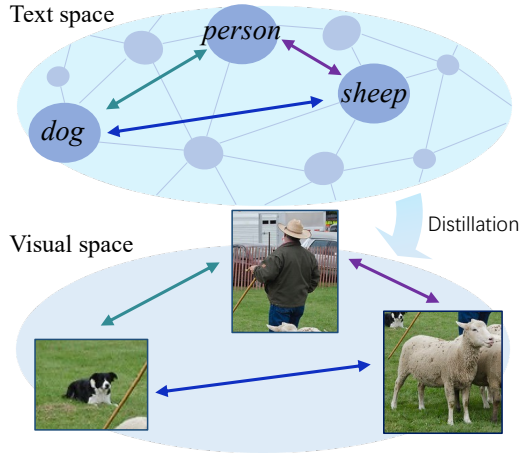
Figure 3. **Text-guided knowledge distillation**. Instead of learning a single visual representation from CLIP, our method utilizes the distance among corresponding categories in text space as guidance to learn a structure of various objects in the visual space.

w_i in the training set. We manually filter out noisy synonyms with semantic ambiguity, such as "rock and roll" for the terrain "rock" category, to obtain a precise synonym set. However, while the generated synonyms can be used to describe the whole category, for a specific instance, there may be a more appropriate word to use. For example, "child" and "man" are both hyponyms of "person", but it is not appropriate to use "child" to describe someone in their 40s. To address this issue, we introduce a new synonym score metric that measures the distance between a synonym word and a visual instance.

During training, we randomly switch the ground truth text prompt for an instance Ins_k with i-th synonym w_k^i from its category's synonym set, using the synonym score as the probability, which is calculated as follows:

$$S_i = \frac{\exp(\mathcal{R}(Ins_k) \cdot \mathcal{T}(w_k^i))}{\Sigma_{j=1}^{N_k} \exp(\mathcal{R}(Ins_k) \cdot \mathcal{T}(w_k^j))} \quad (1)$$

where $\mathcal{R}$ is the CLIP [37] vision encoder, which takes images cropped by instance masks as input, and $\mathcal{T}$ is the CLIP text encoder, which takes a synonym word from the category's synonym set as input. N_k is the size of the synonym set, and $\cdot$ represents the cosine similarity calculation. Our text diversification strategy prevents overfitting to specific words and enriches the text prompts with more varied and meaningful synonyms.

3.2. Text-Guided Knowledge Distillation

The pre-trained CLIP model is crucial for identifying novel classes and achieving cross-modal alignment. A straightforward approach to leverage CLIP is to incorporate a frozen CLIP image encoder to extract visual embeddings

for each mask. Although this approach has shown promising results in recent studies [44, 14], it results in high computation overhead since the CLIP vision encoder must be repeatedly forward passed for each mask proposal, as shown in Tab. 2. Additionally, since the frozen CLIP encoder is not fine-tuned with known categories, it fails to utilize the training priors to improve recognition of seen categories.

We propose leveraging the well-aligned CLIP space and utilizing knowledge distillation to enhance the generalization ability of visual embeddings. During training, we employ a frozen CLIP image encoder as a teacher model. The teacher model takes images masked by ground truth masks as input and generates region-level visual embeddings for each mask. By imposing constraints between the learned visual queries and the corresponding region embeddings produced by the CLIP teacher, we can take advantage of the superior pre-trained weights of CLIP without increasing the inference process's complexity. This vanilla knowledge distillation can be formulated as:

$$\mathcal{L}_{KD} = \frac{1}{N} \sum_{i=1}^{N} \|\mathcal{V}_i - \mathcal{R}(I, M_i)\|, \quad (2)$$

where N is the number of ground truth masks in the image. $\mathcal{V}_i$ denotes the generated visual queries matching i-th ground truth. M_i is the i-th ground truth mask used to mask the image I to serve as input for the CLIP teacher. $\|A - B\|$ denotes the distance measure between A and B.

Although the aforementioned vanilla knowledge distillation strategy can fit the representations of individual categories into the CLIP space, it fails to consider the relationships between objects of different categories, making it challenging to build an overall space similar to the pre-trained CLIP. To overcome this limitation, we propose a **text-guided** knowledge distillation strategy that utilizes the regions of all categories present in the image to calibrate the representation space of the trained model. As illustrated in Fig. 3, in a well-aligned CLIP space, the relationship between the visual representations of different categories should be consistent with the relationship between the corresponding texts. Therefore, we can use the distance between category names in the text space as a guidance signal for distilling visual embeddings.

For instance, taking the "bus" and "bear" as examples, when we distill the student features belonging to the "bus" class, the distance between the student visual features and the teacher CLIP features of the "bear" region should be the same as the distance between the text embeddings of "bus" and "bear". The text-guided knowledge distillation process can be formulated as:

$$\mathcal{L}_{TGKD} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \left\| (\|\mathcal{V}_i - \mathcal{R}(I, M_j)\| - \|\mathcal{T}(Y_i) - \mathcal{T}(Y_j)\|) \right\|,$$

$$(3)$$

where $\mathcal{T}$ denotes the CLIP text encoder and Y_i is the category name of the i-th ground truth region.

3.3. Loss Functions

The total loss consists of three parts: segmentation loss $\mathcal{L}_M$, alignment loss $\mathcal{L}_A$, and knowledge distillation loss $\mathcal{L}_{TGKD}$. To supervise the output mask proposals, we adopt a combination of binary cross-entropy loss and dice loss [33], following query-based segmentation methods [8, 7]. For the alignment loss $\mathcal{L}_A$, we utilize cross-entropy to supervise the matching scores. Additionally, we incorporate the grounding loss, following prior work [18, 20, 46], to leverage the image-level captions and encourage region-word alignments. Specifically, the grounding loss maximizes the similarity score of the labeled image-caption pair over all images and all captions in a mini-batch [18]. The total loss is formulated as follows:

$$\mathcal{L} = \lambda_m \mathcal{L}_M + \underbrace{\lambda_c \mathcal{L}_{CE} + \lambda_g \mathcal{L}_G}_{\mathcal{L}_A} + \lambda_{kd} \mathcal{L}_{TGKD}, \quad (4)$$

where λ represents the weight of each loss. $\mathcal{L}_{CE}$ is cross-entropy loss. $\mathcal{L}_G$ denotes the grounding loss.

3.4. Open-Vocabulary Video Segmentation

In order to expand the open-vocabulary task to a broader range of applications, we conduct a preliminary exploration of open-vocabulary video segmentation. Specifically, we divide the large-scale video segmentation dataset VIPSeg [32] into seen and unseen categories and construct a baseline using our method.

We follow the common video task [45, 3, 23, 12, 13] training strategy, which first pre-trains the model on COCO [29] and then finetunes on the video dataset. To prevent category information leakage from COCO pre-training, categories of VIPSeg are progressively verified. To this end, we recruited four participants who were asked to recognize the pattern of VIPSeg categories on COCO samples and split VIPSeg categories into three sets. The first set contains categories that are annotated by both datasets and have the same category definition, while the second set contains categories that are annotated by both datasets but differ in the level of granularity, *e.g.*, "ball net" and "goal" *vs.* "net". The third set includes categories that are either treated as background or not found in COCO samples, such as "tyre". Finally, we select 12 categories from the third set as novel categories, which cover a total of 9 super-categories defined by VIPSeg.

Similar to Video Mask2Former [6], we represent the entire video sequence as a 3D spatio-temporal volume of dimensions $T \times H \times W$, where T is the number of frames, H and W are the height and width, respectively. By extending our approach in a manner similar to Video Mask2Former, our method can be easily adapted to the video scenario.

4. Experiment

Following previous works [44, 18, 14], we evaluate our open-vocabulary image segmentation model in a cross-dataset setting. In this setting, the model is trained on one dataset and evaluated on other datasets without fine-tuning or retraining. The default training dataset in this paper is COCO Panoptic [29] with 133 categories, as used in previous works [18]. This setting is particularly challenging as the model has to handle both unseen classes and domain gaps between different datasets [44].

Open-Vocabulary Video Segmentation Setting. Due to the limited number of large-scale video segmentation datasets, we evaluate our open-vocabulary video segmentation model in the ordinary zero-shot setting. In this setting, the model is trained on the seen categories and evaluated on both seen and unseen categories.

4.1. Datasets and Evaluation Metrics

To evaluate the effectiveness of our method, we conduct extensive experiments on the image and video datasets, COCO [29], ADE20K [50], Cityscapes [9], Pascal VOC 2012 [15], Pascal Context [34], and VIPSeg [32].

COCO is a large-scale dataset with 117k training images and 5k validation images. We use its panoptic and caption annotations during our training stage, and evaluate it in semantic segmentation manner.

ADE20K contains 20k training images, 2k validation images, and 3k testing images. There are two splits of this dataset. ADE20K-150 contains 150 semantic classes whereas ADE20K-857 has 857 classes. In this paper, we take both splits to verify the performance of our method.

Cityscapes is a scene parsing dataset with 5,000 accurately annotated images and 20,000 coarsely annotated images. Following previous works [9, 44], we take 1,525 images of 19 classes in the accurately annotated set for validation.

Pascal VOC 2012 contains 11,185 training images and 1,449 validation images from 20 classes. We use the provided augmented annotations.

Pascal Context is an extension of Pascal VOC 2010, containing 4,998 training images and 5,005 validation images. In this paper, we take the commonly used PC-59 and challenging PC-459 version for validation.

VIPSeg contains 124 classes, including 3,536 videos and 84,750 frames with pixel-level panoptic annotations.

Evaluation Metric. Following previous works [44, 14, 18], we take the *mean-intersection-over-union* (mIoU) as the metric to measure the segmentation performance. For video datasets, we apply the mIoU on seen classes, unseen classes, and their harmonic mean as major metric.

4.2. Implementation Details

Our implementation is based on `detectron2` [40]. All image-based models are trained with batch size of 112 and

Table 1. The open-vocabulary segmentation performance comparison on the popular image segmentation datasets. PAS denotes the Pascal VOC [15] dataset and PC denotes the Pascal Context [34] dataset. The best results of each dataset are **bolded**. The second best results are underlined. The results on the COCO dataset demonstrate the segmentation ability for the training seen categories. Results on other datasets show the open-vocabulary segmentation ability. † denotes the reproduced result using the same training setting for fair comparison.

Model	Backbone	Training Set	COCO	PAS-20	Cityscapes	ADE20K-150	ADE20K-847	PC-59	PC-459
ZS3Net [1]	R-101	PASCAL-15	-	38.3	-	-	-	19.4	-
SPNet [41]	R-101	PASCAL-15	-	18.3	-	-	1.6	24.3	-
LSeg [25]	R-101	PASCAL-15	-	47.4	-	-	-	-	-
ZegFormer [14]	R50	COCO Stuff	-	<u>80.7</u>	-	16.4	-	-	-
Simbaseline [44]	CLIP R-101	COCO Stuff	-	74.5	-	15.3	-	-	-
LSeg+ [18]	R-101	COCO Panoptic	-	59.0	-	13.0	2.5	36.0	5.2
OpenSeg [18]	R-101	COCO Panoptic	36.9	60.0	-	15.3	**4.0**	36.9	6.5
Simbaseline† [44]	CLIP R-50	COCO Panoptic	39.5	-	30.0	14.4	-	40.1	<u>6.7</u>
Ours	CLIP R-50	COCO Panoptic	<u>49.8</u>	78.7	<u>34.3</u>	<u>17.5</u>	3.2	<u>41.9</u>	6.5
Ours	R-101	COCO Panoptic	**51.2**	**83.2**	**34.8**	**18.8**	<u>3.5</u>	**45.2**	**7.1**

training iteration of 50k. The base learning rate is 0.0003 with a step schedule, in which the steps are set to 40k and 45k, the scaling factor is 0.1. The input image is resized to 512×512. For data augmentation, random horizontal flip and multi-scale jittering with a random scale between [0.8, 1.2] are applied. For the weights of the loss function, we set λ_m to 5, λ_c, λ_g and λ_{kd} to 2 by default. All video-based models are fine-tuned with batch size of 16 and training iteration number of 3k. The base learning rate is 0.0001 with a step schedule, and the step is set to 2k. The backbones of both our model and the distillation teacher CLIP model are CLIP ResNet-50 [37] by default. Note that the text encoder of CLIP is frozen in the training stage. Other hyperparameters are the same as Mask2Former [7].

4.3. Comparison with State-of-the-Art Methods

We evaluate the effectiveness of our proposed method against state-of-the-art techniques on several popular image segmentation datasets [29, 50, 34, 15, 9], to assess its open-vocabulary performance. The results are presented in Tab. 1. The obtained results indicate that our method demonstrates strong open-vocabulary segmentation ability. Specifically, when trained on the COCO Panoptic [29] dataset, which contains 133 categories, our method achieves 7.1 mIoU on the complete Pascal Context [34] dataset with 459 categories and 18.8 mIoU on the ADE20K [50] dataset with 150 categories. Moreover, our approach, which utilizes CLIP ResNet-50 as the backbone, outperforms prior work utilizing CLIP ResNet-101. The comparison on COCO [29] dataset verifies the effectiveness of our method for in-domain segmentation tasks. Compared with previous approaches, our method shows remarkable open-vocabulary segmentation capability while significantly improving the recognition of training categories. Additionally, we perform experiments with ImageNet [10] pre-trained backbone, and our model also achieves promising results, demonstrating

Table 2. Computational cost comparison between our method and current two-stage methods. The FLOPs and Params are measured on the backbone of ResNet101. The FPS is recorded on the same single V100 GPU.

Model	FLOPs	Params	FPS
Simbaseline [44]	1165.07G	89.76M	2.32
ZegFormer [14]	1127.86G	63.90M	5.39
Ours	151.44G	40.51M	8.04

the flexibility of our approach.

We also provide a comparison of the computational complexity and efficiency of our method with two previous two-stage methods [14, 44]. As shown in Table 2, existing region-level alignment methods require an additional frozen CLIP [37] vision encoder to extract foreground visual features for each mask proposal, leading to massive models and slower inference speeds. In contrast, our method achieves high segmentation performance while maintaining a reasonable computation cost. The table shows that our method has approximately 10% of the FLOPs of the previous methods, and a significant increase in FPS can also be observed.

4.4. Open-Vocabulary Video Segmentation

Recognizing novel categories in videos is a challenging task due to the complexity and variability of video scenes. The segmentation results of our proposed method on the VIPSeg [32] dataset are presented in 3. Following Video Mask2Former [6], we extend our method to a video version and finetune it on the seen categories. The baseline in 3 refers to the model trained with mask loss, cross-entropy loss, and dice loss [33] only. The corresponding mIoU values of the seen, unseen, and harmonic categories are 44.2, 2.4, and 4.5, respectively. With the addition of text-guided knowledge distillation supervision, the mIoU values of the unseen and harmonic categories improved by 0.5 and 0.9,

Table 3. The quantitative results of the video open-vocabulary segmentation. The model is trained on the seen categories and evaluated on both the seen and unseen categories.

Model	Seen	Unseen	Harmonic
Baseline	44.2	2.4	4.5
+ KD	43.4	2.9	5.4
+ KD + TD	45.8	8.5	14.4

Table 4. The ablation study on the proposed component. TD denotes the text diversification strategy. TGKD is the text-guided knowledge distillation strategy.

TD	TGKD	Pascal Context	Cityscapes
		39.70	27.61
✓		41.45	32.16
	✓	41.23	32.62
✓	✓	41.91	34.35

Table 5. Experiment results of different distillation strategies. Here × indicates that no knowledge distillation is performed. Vanilla denotes the distillation guided by visual features from one ground truth region. Vision-guided means taking visual embeddings from all ground truth regions as supervision.

Distillation Strategy	Pascal Context	Cityscapes
×	39.70	27.61
Vanilla	40.14	32.49
Vision-guided	39.67	32.33
Text-guided	41.23	32.62

Table 6. Experiments of different teacher and student embeddings.

Teacher	Student	Pascal Context	Cityscapes
global token	post	40.71	30.43
global token	prior	40.31	31.13
spatial token	post	41.29	30.12
spatial token	prior	41.23	32.62

Table 7. Experiment results of applying TD to other methods.

Model	TD	Pascal Context	Cityscape
Simbaseline		40.1	30.0
Simbaseline	✓	42.6	34.7

respectively. Moreover, by utilizing our proposed text diversification strategy, the model is able to achieve 8.5 and 14.4 on unseen and harmonic mIoU, respectively, which is almost three times improvement over the baseline method.

4.5. Ablation Study

In this section, we conduct several ablations to justify the design choices in our proposed network.

Component Analysis. To verify the effectiveness of our proposed strategies, we conduct experiments on the Pascal Context [34] and Cityscapes [9] datasets. The results are shown in Tab. 4. In the table, TD denotes the text diversification training strategy, and TGKD denotes the proposed text-guided knowledge distillation. The model is trained on the COCO Panoptic dataset [29] with CLIP ResNet-50 as the backbone. As can be seen from the results, the text diversification strategy improves the performance by about 2% and 4% on Pascal Context and Cityscapes, respectively. The text-guided knowledge distillation also contributes to a performance gain of 1.6% and 4% on these datasets. When both strategies are used together, the final performance is boosted to 41.91% and 34.35%, respectively.

Distillation Methods. We compared different knowledge distillation methods in Tab. 5. For each generated region query, vanilla distillation constrains it using only the CLIP

visual features of its corresponding ground truth region. This approach does not take into account the relationship between the region query and embeddings of other categories, which may compromise the effect of multi-modal alignment. To alleviate this problem, we propose to leverage all regions in the image to supervise each visual query. Since regions have different visual content information, only using the distance between visual embeddings as distillation guide may introduce errors. Thanks to CLIP's excellent pre-trained common space, the generated queries can learn high-level semantic information for each category by using the distance between text embeddings of different categories as guidance. Experiment results also prove that using text distance as guidance works best.

Distillation Features. There are various options of teacher embeddings and student embeddings for knowledge distillation. Specifically, the teacher embedding can be the global token or spatial tokens with mask-based pooling in the attention pooling process of CLIP [37]. For student embedding, we have experimented with the visual queries before and after the projection layer. The results are shown in Tab. 6. We find that the best choice is to use spatial tokens with mask-based pooling as the teacher embedding and the queries before the projection layer as the student embedding.

Text Diversification Strategy. We experiment with different text diversification strategies in Tab. 8 (a), including (1) randomly replacing the GT with synonym with probability being its synonym score described in Sec. 3.1, (2) taking the maximum (GroupMax) or (3) the average among the complete synonym set as the prediction of the corresponding category. All methods are equally effective, showing that text diversification method is robust to different strategies. Technically, our proposed text diversification is general and applicable to other open-vocabulary segmentation methods. We additionally verify its effectiveness by applying to Simbaseline [44], Tab. 7 shows that it improves by 2.5% and 4.7% mIoU on Pascal Context and Cityscapes, respectively.

Different Visual Backbones. We also experiment with different visual backbones for our method. As Tab. 8 shows, the CLIP pre-trained backbones perform better due to the
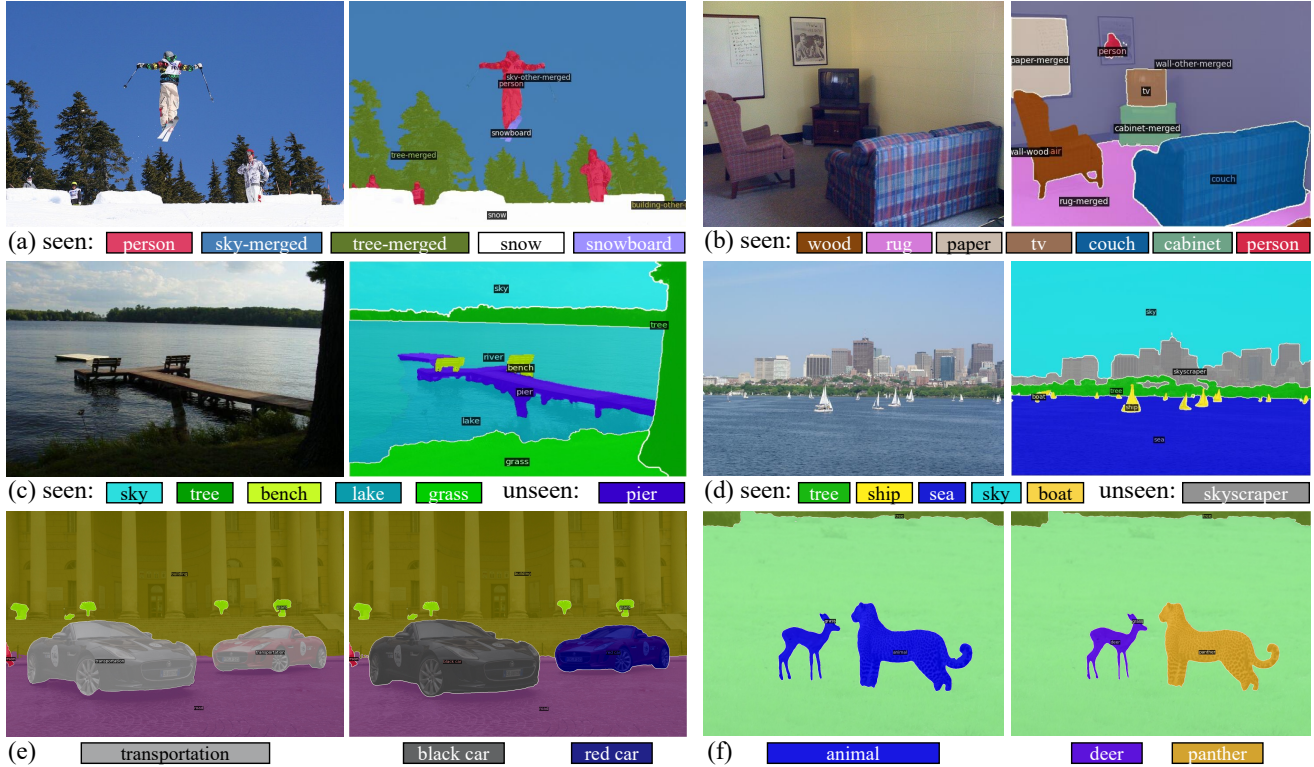
Figure 4. **Qualitative results.** (a) and (b) are evaluation results of COCO panoptic [29] dataset, (c) and (d) are evaluation results of ADE20k-150 [50] dataset, (e) and (f) are inference results of different designed text prompts. For (a)-(d), categories of prediction are shown below, for (e) and (f), difference between twice text prompt inputs are shown below.

Table 8. Experiment results of different text diversification methods and backbones.

	Pascal Context	Cityscapes
(a) Different Text Diversification Method		
Random	41.45	32.16
GroupAvg	41.80	31.42
GroupMax	41.17	32.73
(b) Different Visual Backbones		
ImageNet-R50	45.6	32.9
CLIP-R50	41.9	34.3
ImageNet-R101	45.2	34.8
CLIP-R101	44.2	37.6

Figure 5. **Failure cases** from the ADE20k-150 dataset. Zoom in to check classification results.

well-aligned multi-modal space. However, with the proposed text diversification and text-guided knowledge distillation strategies, our method with ImageNet [10] pretrained backbones performs equally well. This greatly expands the flexibility of our method since we are no longer constrained to vision-language pre-trained backbones.

4.6. Qualitative Results

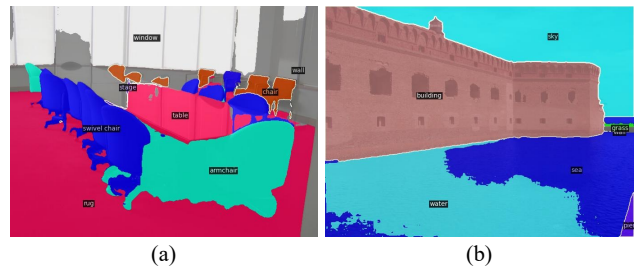Fig. 4 shows some visualization results of our method. From (c) and (d), we can see that our method is able to distinguish the regions of novel categories, *e.g.*, "pier" and "skyscraper", from base categories. As one object can be described differently by multiple descriptions, for the same image, we also tested with different prompts to verify the open-vocabulary segmentation ability of our method. As shown in (e) and (f), our method can distinguish concepts of different granularities (*e.g.*, "transportation" *vs*. "black car" or "red car", "animal" *vs*. "deer" or "panther"). Note that none of these categories are used in COCO Panoptic [29] and our text diversification strategy.

4.7. Failure cases analysis

We investigate the failure cases and make the following observations. When more fine-grained and similar semantic categories are given together, the model usually mixes them up because they share lots of semantic information, *e.g.*, there are "chair", "armchair", "seat" and "swivel chair" labels annotated in ADE20k-150 dataset, while COCO Panoptic labels all of them as "chair".

In the Fig. 5, there are many swivel-able "armchairs" in the middle of (a), the model wrongly segments them as "armchair", "swivel chair" and "chair" separately. In (b), the model confuses the water and sea.

5. Conclusion

In this paper, we propose Global Knowledge Calibration that preserves the generalization ability during the training stage while enabling fast open-vocabulary image segmentation by abandoning the additional frozen CLIP during the inference stage. To broaden the text diversity, we leverage WordNet [17] to avoid collapsing into particular known category names. We also propose text-guided knowledge distillation to utilize the well-aligned multi-modal space of CLIP [37]. Extensive experiments on popular segmentation datasets demonstrate that our method outperforms previous methods in terms of performance and inference cost. To the best of our knowledge, we are the first to explore video open-vocabulary segmentation.

Limitations and Future Work. We notice that our video open-vocabulary segmentation model still suffers from overfitting if we naïvely increase the number training iterations, resulting in performance degradation on novel categories. Despite being significantly faster than existing two-stage methods, the running speed is still far from real-time. We will study these in future work.

References

[1] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. In *NeurIPS*, 2019. 2, 6

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3

[3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2018. 1

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019. 2

[6] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G. Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 5, 6

[7] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 5, 6

[8] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 3, 5

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 5, 6, 7

[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*, 2009. 6, 8

[11] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *CVPR*, 2018. 1

[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. 5

[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. 5

[14] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6

[15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 5, 6

[16] Yan Fang, Feng Zhu, Bowen Cheng, Luoqi Liu, Yunchao Wei, and Yao Zhao. Locating noise is halfway denoising for semi-supervised segmentatio. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1

[17] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998. 2, 3, 9

[18] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Open-vocabulary image segmentation. *arXiv preprint arXiv:2112.12143*, 2021. 1, 2, 3, 5, 6

[19] Meng-Hao Guo, Chengze Lu, Qibin Hou, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *arXiv preprint arXiv:2209.08575*, 2022. 1

[20] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *ECCV*, 2020. 5

[21] Shuting He, Henghui Ding, and Wei Jiang. Primitive generation and semantic-related alignment for universal zero-shot segmentation. In *CVPR*, 2023. 1

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

[23] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. 2022. 5

[24] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 2

[25] Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. 1, 2, 6

[26] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, 2020. 2

[27] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 2

[28] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted CLIP. *arXiv preprint arXiv:2210.04150*, 2022. 2

[29] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014. 5, 6, 7, 8

[30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[31] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 2

[32] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *CVPR*, 2022. 2, 5, 6

[33] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016*, 2016. 5, 6

[34] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan L. Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014. 5, 6, 7

[35] Mengxue Qu, Yu Wu, Wu Liu, Qiqi Gong, Xiaodan Liang, Olga Russakovsky, Yao Zhao, and Yunchao Wei. Siri: A simple selective retraining mechanism for transformer-based visual grounding. In *European Conference on Computer Vision*. Springer, 2022. 2

[36] Mengxue Qu, Yu Wu, Yunchao Wei, Wu Liu, Xiaodan Liang, and Yao Zhao. Learning to segment every referring object point by point. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3, 4, 6, 7, 9

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 1

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 3

[40] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5

[41] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero- and few-label semantic segmentation. In *CVPR*, 2019. 1, 2, 6

[42] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NIPS*, 2021. 1

[43] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas M. Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, 2022. 2

[44] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *arXiv preprint arXiv:2112.14757*, 2021. 1, 2, 3, 4, 5, 6, 7

[45] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. *CoRR*, abs/1905.04804, 2019. 5

[46] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021. 5

[47] Hui Zhang and Henghui Ding. Prototypical matching and open set rejection for zero-shot semantic segmentation. In *ICCV*, 2021. 1, 2

[48] Zekang Zhang, Guangyu Gao, Zhiyuan Fang, Jianbo Jiao, and Yunchao Wei. Mining unseen classes via regional objectness: A simple baseline for incremental segmentation. *NeurIPS*, 35, 2022. 1

[49] Zekang Zhang, Guangyu Gao, Jianbo Jiao, Chi Harold Liu, and Yunchao Wei. Coinseg: Contrast inter- and intra- class representations for incremental segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1

[50] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, 2017. 5, 6, 8