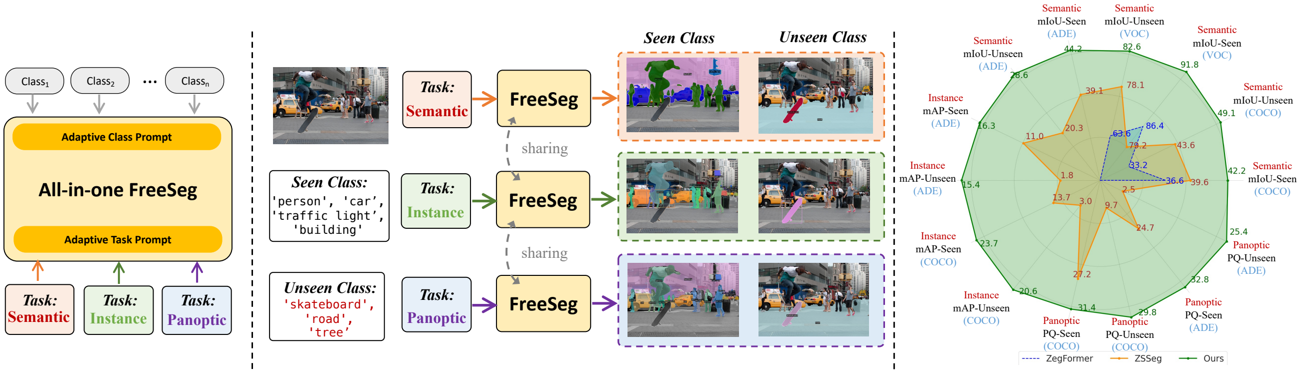
FreeSeg: Unified, Universal and Open-Vocabulary Image Segmentation

Jie Qin[1,2,3]⋆ Jie Wu[2]⋆ Pengxiang Yan[2] Ming Li[2] Ren Yuxi[2] Xuefeng Xiao[2]
Yitong Wang[2] Rui Wang[2] Shilei Wen[2] Xin Pan[2] Xingang Wang[1]†
[1]Institute of Automation, Chinese Academy of Sciences [2]ByteDance Inc
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences



(a) One-shot Training *(b) Multi-task Inference* *(c) Performance Diagram*

Figure 1. We propose FreeSeg, a generic framework to accomplish Unified, Universal and Open-Vocabulary Image Segmentation. (a) FreeSeg optimizes an all-in-one network via one-shot training. (b) FreeSeg employs the same architecture and parameters to handle diverse segmentation tasks seamlessly in the inference procedure. (c) FreeSeg establishes new state-of-the-art performance across diverse segmentation tasks, training datasets and zero-shot generalization.

Abstract

Recently, open-vocabulary learning has emerged to accomplish segmentation for arbitrary categories of text-based descriptions, which popularizes the segmentation system to more general-purpose application scenarios. However, existing methods devote to designing specialized architectures or parameters for specific segmentation tasks. These customized design paradigms lead to fragmentation between various segmentation tasks, thus hindering the uniformity of segmentation models. Hence in this paper, we propose ***FreeSeg****, a generic framework to accomplish* ***Unified****,* ***Universal*** *and* ***Open-Vocabulary*** *Image Segmentation. FreeSeg optimizes an all-in-one network via one-shot training and employs the same architecture and parameters to handle diverse segmentation tasks seamlessly in the inference procedure. Additionally, adaptive prompt learning facilitates the unified model to capture task-aware and category-sensitive concepts, improving model robustness in multi-task and varied scenarios. Extensive experimental results demonstrate that FreeSeg establishes new state-of-the-art results in performance and generalization on three segmentation tasks, which outperforms the best task-specific architectures by a large margin:* ***5.5%*** *mIoU on semantic segmentation,* ***17.6%*** *mAP on instance segmentation,* ***20.1%*** *PQ on panoptic segmentation for the unseen class on COCO. Project page: https://FreeSeg.github.io.*

1. Introduction

Image segmentation has been one of the most widely researched topics in computer vision, aiming to simultaneously group and categorize object pixels in the image. In the recent literature, the image segmentation community has witnessed tremendous success at cost of large-scale datasets [1, 3, 30], where objects are exhaustively annotated with pixel-level masks and category labels. However, due to the time-consuming and laborious annotations, the template categories sizes of existing segmentation tasks are still limited to an order of 10 or 10^2, which is in orders of magnitude much smaller than the vocabulary that humans use to describe the real world. Such learning objective binds the segmentors' scalability into a limited cognitive space, and it becomes a critical bottleneck when this system is popularized to handle richer and more generalized semantics.

⋆Equal contribution. †Corresponding author. This work was done while Jie Qin interned at ByteDance.

As a viable path to handle categories of custom specification beyond the training dataset, open-vocabulary learning leverages large-scale visual-language pre-training models (such as CLIP [26], ALIGN [14]) to calculate matching similarity between visual concept and text corpus. Recently, a series of segmentation-based open-vocabulary studies [1, 37, 38] have emerged to design task-specific architectures and parameters for individual segmentation task. For example, ZSSeg [38] leverages the off-the-shelf pre-trained CLIP model and achieves competitive performance in open vocabulary semantic segmentation. However, current works suffer from two obvious shortcomings when popularized to general segmentation scenes: i) *task-insensitive*: they can not capture task-aware characteristics and be effectively generalized to diverse segmentation tasks; ii) *resource-unfriendly*: the model needs to be trained from scratch when switching tasks, and diverse tasks require deploying multiple customized models. Although MaskFormer [6] succeeds in accomplishing multiple segmentation tasks into one compact system, it still needs to train a customized model for each task and it is not designed for open-vocabulary tasks. These observations motivate us to raise a question: *how to design a unified open-vocabulary framework to accomplish universal segmentation tasks?*

To address the above question, As shown in Fig.1, we propose **FreeSeg**, a novel framework to accomplish **Unified**, **Universal** and **Open-Vocabulary** Image Segmentation. In FreeSeg, our goals are mainly three-fold: i) Unified: FreeSeg designs a unified (all-in-one) network that employs the same architecture and inference parameters to handle multiple segmentation tasks; ii) Universal: FreeSeg adapts to various tasks, namely semantic, instance and panoptic segmentation; iii) Open-Vocabulary: FreeSeg is capable of generalizing to arbitrary segmentation categories.

In general, FreeSeg advocates a two-stage segmentation framework, with the first stage extracting universal mask proposals and the second stage accomplishing zero-shot classification on these masks. Specifically, FreeSeg conducts a one-shot training procedure to optimize a unified segmentation model with multi-task labels, which helps to capture task-special characteristics for universal segmentation. An adaptive prompt learning scheme is introduced to encode task-aware and category-sensitive concepts into the text abstraction. It enables FreeSeg to flexibly accomplish different segmentation tasks of arbitrary categories, handling all tasks and categories in one model. To sum up, FreeSeg is a *task-flexible, category-arbitrary and performance-excellent* framework, the main contributions of our work are listed as follows:

- To the best of our knowledge, we offer the first attempt to tackle a novel computer vision task, namely, unified open-vocabulary segmentation. A universal framework FreeSeg is proposed to employ an all-in-one model with the same architecture and inference parameters to accomplish open-vocabulary semantic, instance, and panoptic segmentation.
- Adaptive prompt learning explicitly encodes multi-granularity concepts (task, category) into compact textual abstraction and helps the unified model generalize to arbitrary text descriptions. FreeSeg further designs the semantic context interaction and test time prompt tuning mechanism to improve cross-model alignment and generalization for unseen classes.
- We evaluate FreeSeg on three image segmentation tasks (semantic, instance, and panoptic segmentation) using COCO, ADE20K and VOC 2012. As shown in Fig.1 (c), extensive experiments demonstrate that FreeSeg establishes new state-of-the-art results in terms of performance and generalization. In addition to reducing the research effort by at least three times, it outperforms the best-specialized architectures and is more feasible for multi-task deployment.

2. Related Work

2.1. Open Vocabulary Segmentation

Deep learning [18,19,29,34–36,39] and image segmentation has recently witnessed tremendous success [3,4,6,24,25,30,40]. Open vocabulary segmentation aims to segment the target categories that can not access during the training procedure. The existing approaches can be divided into two aspects: mapping visual features into semantic space [1,11,37] and cross-modal alignment with pre-trained models [7, 17, 38]. For the mapping aspect, SPNet [37] encodes visual features to the semantic embedding space and then projects each pixel feature to predict probabilistic outcomes through a fixed semantic word encoding matrix. ZS3Net [1] generates the pixel-level features of unseen classes in the semantic embedding space and adopts the generated features to supervise a visual segmentation model. STRICT [23] introduces a self-training technique into SPNet to improve the segmentation performance of unseen classes. Cross-modal alignment employs robust zero-shot capabilities of the pre-trained cross-modal models such as CLIP [26] to conduct open vocabulary segmentation tasks. LSeg [17] learns a CNN model to compute per-pixel image features to match with the text embeddings embedded by the pre-trained text model. ZegFormer [7] and ZSSeg [38] leverage the visual model to generate the class-agnostic masks, and use the pre-trained text encoder to retrieve the unseen class masks. XPM [13] utilizes the region-level features to match CLIP-based text embeddings to accomplish the open vocabulary instance segmentation. MaskCLIP [8] attempts to establish relationships between the class-agnostic masks in the CLIP visual encoder to complete the open vocabulary panoptic segmentation.

2.2. Universal Segmentation Architecture

The goal of the universal segmentation framework is to employ the same architecture in arbitrary segmentation tasks, so current universal segmentation approaches [5, 6, 41] regularly constrain multiple tasks (*semantic*, *instance*, *panoptic*) to a unified training paradigm. MaskFormer [6] unifies the segmentation tasks into a classification problem for masks, *i.e*., outputting binary masks and the corresponding categories, which achieves state-of-the-art performance in both semantic and panoptic segmentation tasks. K-Net [41] standardizes instance segmentation into semantic segmentation via learnable kernels to accomplish the semantic, instance, and panoptic segmentation tasks simultaneously. Mask2Former [5] employs the masked attention mechanism into MaskFormer to improve the generalization of the unified model and the performance of each task. However, these unified frameworks still require training a separate model for each task to achieve the best performance. Our proposed FreeSeg conduct one-shot training to optimize an all-in-one model to finish multiple segmentation tasks.

2.3. Prompt Learning

Prompt learning achieved a remarkable leap in the field of NLP [12, 16, 33], and then is rapidly popularized into the vision or vision-language models [28, 45]. CoOp [45] brings continuous prompt optimization from downstream data to adapt the pre-trained vision-language model. DenseCLIP [28] finetunes the pre-trained text encoder with the given prompt templates to perform text and visual feature matching for downstream intensive prediction tasks such as detection and segmentation. For open vocabulary segmentation tasks [7, 38], prompt templates are generated from the given category names, and then are encoded to the text embeddings for matching the unseen classes.

3. Methodology

3.1. FreeSeg Framework

The proposed unified open-vocabulary segmentation aims to optimize an all-in-one model to obtain semantic, instance, and panoptic segmentation results on arbitrary categories. To address this novel task, we propose a novel framework to accomplish unified and universal open vocabulary segmentation in this paper, termed as FreeSeg. FreeSeg advocates a two-stage framework, with the first stage extracting universe mask proposals and the second stage leveraging CLIP to perform zero-shot classification on the masks which are generated in the first stage. The whole framework of FreeSeg is illustrated in Fig. 2.

Training. The training data in the first stage contains images I, seen category set C_{seen}, task names T_{train} and multi-task labels M^{gt}. The training procedure only accesses the seen categories C_{seen} and the corresponding labels. The mask proposal extractor encodes the image into visual concepts $F_v \in \mathcal{R}^{N\times D}$ and class-agnostic masks $M \in \mathcal{R}^{N\times H\times W}$, where N and D denote the number of queries and feature dimensions. To encapsulate multiple learned tasks in a unified model, We leverage three task-specific labels, *i.e*., $M^{gt} \in (M^{gt}_{sem}, M^{gt}_{ins}, M^{gt}_{pan})$ to selectively supervise the mask proposal extractor with mask loss:

$$\mathcal{L}_{mask} = \mathcal{L}_F(M, M^{gt}) + \mathcal{L}_D(M, M^{gt}), \tag{1}$$

where $\mathcal{L}_F$ denotes the Focal [20] loss and $\mathcal{L}_D$ is the Dice [22] loss. Simultaneously optimizing all tasks is often difficult due to gradient conflicts across tasks during training, thus only one task label is selected for supervision per iteration, which is randomly selected from the $(M^{gt}_{sem}, M^{gt}_{ins}, M^{gt}_{pan})$.

To facilitate FreeSeg to handle task and categories characteristics, we design a novel adaptive prompt learning to explicitly embed task and category concepts into joint text embeddings $F_t \in \mathcal{R}^{C\times D}$ via a pre-trained CLIP-based text encoder, where C denotes the number of categories. The cross-modal classification supervision is set up to enable FreeSeg to classify generated masks according to arbitrary text. Specifically, the visual concepts F_v are leveraged to compute the similarity matching map with text embeddings F_t. The cosine similarity score $\mathcal{S} \in \mathcal{R}^{N\times C}$ between pairs of F^i_v and F^j_t is computed as:

$$S(i,j) = cos(F^i_v, F^j_t) = \frac{F^i_v \cdot F^j_t}{\|F^i_v\| \left\|F^j_t\right\|}, \tag{2}$$

where $i \in [1, N]$, $j \in [1, C]$. The obtained similarity matching map indicates the probability of the predicted category for all class-agnostic masks, which is supervised by the class labels with the cross-entropy loss $\mathcal{L}_{cla}$. The total training loss is formulated as:

$$\mathcal{L} = \mathcal{L}_{cla} + \mathcal{L}_{mask}, \tag{3}$$

Testing. In the testing phase, the trained mask proposal extractor generates a set of binary masks with textual guidance and leverages the pre-trained CLIP visual encoder to obtain mask-level visual concepts. FreeSeg calculates the similarity between mask representation and compact text embedding and outputs task-oriented segmentation results according to the adaptive task prompt. With the aid of adaptive prompt learning, FreeSeg can handle arbitrary tasks and categories. The test category set C_{test} consists of seen classes C_{seen} and additional unseen classes C_{unseen}.

3.2. Adaptive Prompt Learning

To encode arbitrary tasks and categories into compact textual abstraction, we propose the adaptive prompt learn-

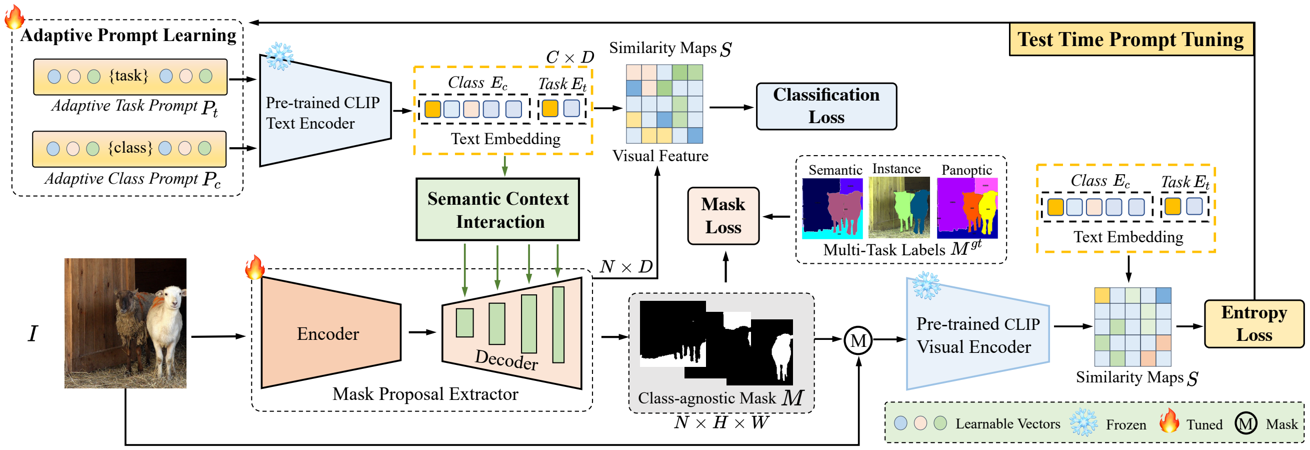


Figure 2. Overview of our two-stage FreeSeg framework. i) one-shot training: optimizes an all-in-one segmentation model via multi-task supervision to generate universal mask proposals; ii) Multi-task inference: leverages pre-trained CLIP to classify mask proposals according to adaptive task and class prompt.

ing module containing the adaptive task prompt P_t and the adaptive class prompt P_c. Fixed prompt puts all category and task names into the same templates, which is not the optimal representation for task-category pair contexts. While adaptive prompt learning turns the task and category texts into a set of learnable vectors, which are concatenated as text embeddings to facilitate model training.

Adaptive Task Prompt. The adaptive task prompt promotes capturing task-specific characteristics, encapsulating multiple learned tasks in a unified framework, and effectively disentangles the parameter spaces to avoid different tasks' training conflicts. Specifically, the adaptive task prompt P_t is generated according to the template $\{\circ \circ ... \; t ... \circ \circ\}$, where $\circ$ denotes the learnable vectors. t is the corresponding task name in a task set T, which contains "*semantic segmentation.*", "*instance segmentation.*", or "*panoptic segmentation.*". Then the task prompts are embedded by the pre-trained CLIP text encoder Ψ:

$$E_t = \Psi(P_t(t)), t \in T, \tag{4}$$

where E_t denotes the task embeddings.

Adaptive Class Prompt. An adaptive class prompt is introduced to popularize FreeSeg to generalize to broader unseen categories and improve open-domain performance. Given the semantic categories C_{seen} involved in training, the class prompts P_c are obtained by the template $\{\circ \circ ... \; c ... \; \circ \circ\}$, where c is the filled class names. The adaptive class prompt P_c is embedded to generate the class text embeddings E_c:

$$E_c = \Psi(P_c(c)), c \in C_{seen}, \tag{5}$$

To model a joint task-category textual space, the class text embeddings E_c and the task text embeddings E_t are fused to get the multi-granularity embeddings F_t:

$$F_t = Cat(E_c, E_t), \tag{6}$$

where Cat denotes the concatenation operation. It is worth noting that the input category can be arbitrary, so F_t can seamlessly adapt to unseen categories for open vocabulary segmentation.

3.3. Semantic Context Interaction

The vanilla visual concepts ignore task and category information that can provide more reliable cues for comprehensive inference. To address this issue, we creatively introduce a *semantic context interaction module* to improve the cross-modal feature matching and alignment by effectively aggregating adaptive textual embedding into visual concepts. Specifically, the semantic context interaction module employs the cross-attention module to model the correlations between text embeddings and multiple-scale visual features.

$$Attn(Q^z, K, V) = softmax(\frac{Q^z K^T}{\sqrt{d_k}})V^T, \tag{7}$$

$$Q^z = \phi_q(F_v^z), \quad K = \phi_k(F_t), \quad V = \phi_v(F_t), \tag{8}$$

where F_v^z denotes z-layer visual feature from decoder in mask proposal extractor. Q^z, K, V denote the query, key, and value embeddings generated by the projection layers ϕ_q, ϕ_k, ϕ_v. $\sqrt{d_k}$ represents the scaling factor. Then the attention relationship is utilized to enhance the visual features:

$$\hat{F_v^z} = \mathcal{H}\{Attn[\phi_q(F_v^z), \phi_k(F_t), \phi_v(F_t)]\}, \tag{9}$$

where $\mathcal{H}$ denotes the output projection layer. The enhanced visual feature $\hat{F_v^z}$ is beneficial to emphasize the visual feature concerning the given text classes.

Table 1. Comparison with state-of-the-art methods in open vocabulary semantic segmentation. mIoU^s and mIoU^u denote the mIoU(%) of seen classes and unseen classes. The variant "Full Sup." denotes training FreeSeg with all seen and unseen classes.

Method	COCO			VOC2012			ADE20K		
	mIoU^s	**mIoU^u**	**hIoU**	**mIoU^s**	**mIoU^u**	**hIoU**	**mIoU^s**	**mIoU^u**	**hIoU**
Full Sup.	42.9	54.3	47.9	92.3	89.5	91.1	46.1	41.5	44.0
SPNet [37]	34.6	26.9	30.3	77.8	25.8	38.8	-	-	-
ZS5 [1]	34.9	10.6	16.2	78.0	21.2	33.3	-	-	-
CaGNet [11]	35.6	13.4	19.5	78.6	30.3	43.7	-	-	-
STRICT [23]	35.3	30.3	32.6	82.7	35.6	73.3	-	-	-
ZegFormer [7]	36.6	33.2	34.8	86.4	63.6	73.3	-	-	-
ZSSeg [38]	39.6	43.6	41.5	79.2	78.1	79.3	39.1	20.3	31.6
Ours	**42.2**	**49.1**	**45.3**	**91.8**	**82.6**	**86.9**	**44.2**	**28.6**	**39.8**

3.4. Test Time Prompt Tuning

To improve the cross-modal alignment of unseen categories, we leverage the test time adaptation (TTA) algorithm [15, 31, 32] to refine the adaptive class prompt during testing, termed as *Test Time Prompt Tuning*.

In the testing phase, we filter out the cosine similarity scores S_u of unseen classes and calculate the corresponding entropy:

$$entro = -\frac{1}{N_u}\sum_{i=1}^{N_u} s_i log(s_i), \tag{10}$$

where $entro$ denotes the entropy value of each sample. N_u is the number of the unseen classes and s_i is the score of i^{th} class of S_u. Then we select the high-confidence queries according to the entropy $S_u^* = S_u[entro < \tau]$, where τ is the threshold of the high confidence. Because the low entropy value indicates the high confidence level of the sample predictions. We calculate the entropy loss $\mathcal{L}_{ent}$ to optimize the parameters of the adaptive class prompt:

$$\mathcal{L}_{ent} = -\frac{1}{N_u K}\sum_{i=1}^{N_u}\sum_{j=1}^{K} s_{ij} log(s_{ij}), \tag{11}$$

where s_{ij} denotes the score of j-th selected queries. K is the queries number in S_u^*.

4. Experimental Results

4.1. Datasets and Evaluation Metrics

4.1.1 Datasets

COCO. COCO dataset [21] contains multi-tasks ground-truth labels towards the same image. We collect semantic labels of COCO stuff [2] and panoptic labels of COCO and merge them to get the unified, category-wide annotations M^{gt}. We follow [37, 38] to divide all 171 categories into 156 seen and 15 unseen classes to complete the open vocabulary segmentation task.

ADE20K. ADE20K [44] contains 20,000 training images and 2,000 validation images with 150 categories. We split 15 categories into unseen classes, and the remaining 135 are treated as seen/training classes.

PASCAL VOC2012. We conduct experiments on PASCAL VOC2012 [9] to accomplish semantic segmentation. Following [37, 38], we divide 20 foreground classes into 15 seen classes and 5 unseen classes to evaluate the effectiveness of the open vocabulary segmentation.

4.1.2 Evaluation Metrics

Semantic segmentation. We follow [7, 38] to adopt the mean of Interaction over Union (mIoU) to respectively evaluate the open vocabulary semantic segmentation performance for seen and unseen classes. We also employ the harmonic mean IoU (hIoU) among the seen and unseen classes to measure comprehensive performance.

Instance segmentation. We report the mean Average Prediction (mAP) of seen and unseen classes for open vocabulary instance segmentation.

Panoptic segmentation. For open vocabulary panoptic segmentation, we follow the setting of fully supervised panoptic segmentation and use task-aware metrics (*i.e.*, PQ, SQ, RQ) to evaluate panoptic segmentation quality.

4.2. Implementation Details

COCO. We employ Mask2Former [5] as the mask proposal extractor and ResNet101 as the backbone. VIT-B/16 is adopted as the backbone of CLIP [26]. All experiments are conducted on 8×A100 GPUs. We take the batch size of 32 per GPU and set the input image size as 640×640. The optimizer is AdamW with a learning rate of 0.0002 and weight decay of 0.0002. The number of training iterations is 60,000. In addition, the learnable parameter size of the task prompt is 8×512, and the class prompt is 16×512. We follow the comparison methods [7, 23, 38] to employ the self-training technique for training.

ADE20K and PASCAL VOC2012. ADE20K dataset uses

Table 2. Comparison with state-of-the-art methods in open vocabulary instance segmentation. mAP^s and mAP^u denote the mAP(%) results of seen classes and unseen classes.

Method	**mAP^s**	**mAP^u**	**mAP**	**AP^{50}**	**AP^{75}**
			COCO		
Full Sup.	24.9	25.1	24.9	37.8	25.8
CLIP [26]	8.5	2.6	7.9	11.8	7.5
ZSSeg [38]	13.7	3.0	12.8	20.9	13.3
PL [27]	34.0	12.4	-	-	-
BLC [42]	36.0	13.1	-	-	-
ZSI [43]	**38.7**	13.6	-	-	-
Ours	23.7	**20.6**	**22.8**	**36.0**	**24.1**
			ADE20K		
Full Sup.	20.3	18.1	20.1	31.4	21.1
CLIP [26]	5.6	3.5	5.4	7.7	5.8
ZSSeg [38]	11.0	1.8	9.8	17.8	9.5
Ours	**16.3**	**15.4**	**16.2**	**25.3**	**16.9**

Table 3. Comparison with state-of-the-art methods in PQ(%), SQ(%), RQ(%) on open vocabulary panoptic segmentation.

Method		Seen			Unseen	
	PQ	**SQ**	**RQ**	**PQ**	**SQ**	**RQ**
			COCO			
Full Sup.	33.1	78.5	39.5	34.1	80.7	41.5
CLIP [26]	14.3	71.5	18.4	9.2	70.3	11.6
ZSSeg [38]	27.2	76.1	34.7	9.7	71.7	12.2
Ours	**31.4**	**78.3**	**38.9**	**29.8**	**79.2**	**37.6**
			ADE20K			
Full Sup.	35.3	78.1	43.9	30.6	74.2	37.5
CLIP [26]	9.5	62.8	12.1	3.4	61.1	4.7
ZSSeg [38]	24.7	70.7	32.2	2.5	65.1	6.3
Ours	**32.8**	**78.2**	**40.4**	**25.4**	**75.2**	**30.6**

512×512 input image size and the number of iterations is set to 20,000 on PASCAL VOC2012. The remaining training settings on these two datasets are the same as COCO.

4.3. Comparison to State-of-the-art Methods

Open Vocabulary Semantic Segmentation We compare FreeSeg with current state-of-the-art open vocabulary semantic segmentation methods in Tab.1, including SPNet [37], ZS5 [1], CaGNet [11], STRICT [23], ZegFormer [7], ZSSeg [38]. Tab.1 can be summarized as the following observations: i) FreeSeg achieves 49.1% and 28.6% mIoU towards unseen classes on COCO and ADE20K, which surpasses the previous best method ZSSeg by +5.5% and +8.3%, respectively. It indicates that FreeSeg can adapt to more generalized scenarios. ii) We also report the result of the fully supervised baseline, denoted as "Full Sup.", which is trained on both seen and unseen classes. Remarkably, FreeSeg is only 0.7% and 5.2% worse than the fully supervised baseline "Full Sup." in seen and unseen classes on COCO, respectively. iii) To compare with competitive methods that are only trained on VOC benchmark, we also report the result of FreeSeg in the same setting as previous work. The experimental results show that FreeSeg obtains 91.8%/82.6% mIoU on the seen and unseen classes, which outperforms ZSSeg by 12.6%/4.5%. It further proves that FreeSeg is both robust and excellent for handling multi-tasks and single task.

Open Vocabulary Instance Segmentation As shown in Tab.2, we compare the open vocabulary instance segmentation performance on COCO and ADE20K datasets, including ZSSeg [38], PL [27], BLC [42], and ZSI [43]. Since ZSSeg did not report the result on this task, we reproduce the results by training on the instance segmentation labels with the official code. The variant "CLIP" denotes the direct matching results with the pre-trained CLIP [26] text and visual encoder. FreeSeg achieves 20.6% mAP of unseen classes on COCO, which outperforms the best-performance method ZSI by +7.0% mAP. However, the mAP of the seen classes of ZSI is higher than FreeSeg. It is because ZSI [43] uses box-level supervision, which is more favorable for instance segmentation, while FreeSeg uses more general mask supervision for various segmentation tasks. In addition to COCO, FreeSeg also achieves promising results on ADE20k. For example, FreeSeg achieves 16.3% / 15.4% mAP on seen / unseen classes, which outperforms the baseline CLIP by +10.7% and +11.9% mAP.

Open Vocabulary Panoptic Segmentation Since few works study open vocabulary panoptic segmentation, we report the results of FreeSeg and the CLIP [26] baseline in Tab. 3. We also re-implement ZSSeg [38] on panoptic segmentation labels to accomplish this task. We observe that FreeSeg achieves 29.8% PQ, 79.2% SQ, and 37.6% RQ of the unseen classes, outperforming ZSSeg by 20.1%, 7.5%, and 25.4%, respectively. The main performance improvement comes from the unseen classes, indicating that this semantic segmentation-oriented method like ZSSeg is hard to generalize to other tasks, while our FreeSeg has noticeable generalization capability. On the ADE20K dataset, FreeSeg also achieves the best results with 25.4%, 75.2%, and 30.6% of unseen classes on PQ, SQ, and RQ, respectively. These above multi-task results prove the generalization ability of FreeSeg for unified open vocabulary segmentation tasks.

4.4. Generalization Analysis

We evaluate the generalization of FreeSeg across datasets. Namely, we train FreeSeg on COCO (G_{coco}) or ADE20K (G_{ade}) and directly test it on other datasets without finetuning. This cross-dataset evaluation is non-trivial because of the significant differences in the data distribution and domains. In this setting, all classes of the target dataset are regarded as unseen classes, and we report the segmentation performance of all unseen classes for three

Table 4. Ablation studies of the proposed modules on COCO datasets.

Adaptive Prompt		Context	Prompt	Semantic		Instance		Panopnic					
Class	Task	Interaction	Tuning	mIoU^s	mIoU^u	mAP^s	mAP^u	PQ^s	SQ^s	RQ^s	PQ^u	SQ^u	RQ^u
✗	✗	✗	✗	38.4	4.9	19.2	0.7	25.9	70.4	32.0	0.1	0.2	0.1
✓	✗	✗	✗	39.0	38.5	20.1	8.8	26.2	70.7	32.5	12.0	62.6	15.3
✓	✓	✗	✗	40.8	41.3	22.2	11.8	28.5	74.3	35.8	15.7	67.4	19.8
✓	✓	✓	✗	42.1	42.6	23.7	13.9	30.0	76.5	37.3	18.1	70.5	23.2
✓	✓	✓	✓	**41.9**	**43.3**	**23.9**	**14.6**	**30.4**	**76.7**	**38.1**	**19.2**	**71.4**	**24.1**

Table 5. Generalization performance (in%) of the open vocabulary segmentation on cross datasets.

Method	mIoU	mAP	PQ	SQ	RQ
		COCO → ADE20K			
CLIP [26]	13.8	3.9	8.2	53.1	10.5
Lseg+ [10]	13.0	-	-	-	-
OpenSeg [10]	15.3	-	-	-	-
ZSSeg [38]	16.4	4.0	9.3	58.0	12.2
MaskCLIP [8]	23.7	5.9	15.1	70.4	19.2
Ours	**24.6**	**6.5**	**16.3**	**71.8**	**21.6**
		ADE20K → COCO			
CLIP [26]	14.7	2.7	8.1	66.3	11.0
ZSSeg [38]	17.7	4.3	11.2	66.5	14.9
Ours	**21.7**	**6.6**	**16.5**	**72.0**	**21.6**

Table 6. Generalization performance (in%) of the open vocabulary semantic segmentation on VOC2012 datasets.

Method	mIoU	
	COCO → VOC2012	*ADE20K → VOC2012*
CLIP [26]	71.6	67.1
ZSSeg [38]	82.1	69.2
Ours	**91.9**	**80.1**

segmentation tasks. As shown in Tab. 5, FreeSeg achieves 24.6% mIoU semantic segmentation, 6.5% mAP instance segmentation, 16.3% PQ, 71.8% SQ, 21.6% RQ panoptic segmentation results on ADE20K with G_{coco}, which outperforms the SOTA method MaskCLIP [8] with 0.9% mIoU, 0.6% mAP, 1.2% PQ, and 2.4% RQ, respectively. FreeSeg also obtains the best performance when validating G_{ade} on COCO datasets, achieving 21.7% mIoU, 6.6% mAP, 16.5% PQ, 72.0% SQ, and 21.6% RQ for semantic, instance, and panoptic segmentation, respectively. The generalization results on VOC2012 with G_{coco} and G_{ade} also verify the transferability of FreeSeg in Tab.6

4.5. Ablation Study

Component Analysis We conduct ablation studies to analyze essential components of FreeSeg on COCO datasets in Tab.4. Note that the self-training technique is not applied in these ablations. The primary vision model achieves an inferior performance of 4.9% mIoU and 0.7% mAP on the unseen classes without any text guidance. By introducing the adaptive class prompt, the performance is improved significantly on COCO, especially for the unseen classes. Then the adaptive task prompt and semantic context interaction module is gradually inserted into the framework, which brings out the performance improvement of 2.8% and 1.3% mIoU on COCO dataset, respectively. Furthermore, the experimental results show that test time prompt tuning also improves the unseen classes' performance during inference.

We also explore the effectiveness of the proposed modules on the open vocabulary instance and panoptic segmentation and obtain a highly consistent conclusion with semantic segmentation. It demonstrates that adaptive prompt learning promotes FreeSeg to capture task-aware and category-sensitive characteristics. The semantic context interaction and test-time prompt-tuning help to improve the cross-modal alignment of visual and text features.

Multi-Task Analysis. To validate the advantages of multi-task learning in FreeSeg, we compare the results of the unified multi-task training with the single-task training for specific tasks. As shown in Tab.7, all the results from the multi-task row are obtained from one unified model, while the single-task results are from three individual models. All results are obtained without the self-training technique. Multi-task training achieves 41.9% and 43.3% mIoU for the seen and unseen classes on open vocabulary semantic segmentation, suppressing the performance of the single-task model. Open-vocabulary instance and panoptic segmentation also show consistent results as semantic segmentation, especially in the performance of unseen classes. FreeSeg improves all metrics of unseen classes on all tasks, proving that the multi-task training scheme can efficiently improve the generalization of the networks. Furthermore, the unified open vocabulary model conducts a one-shot training procedure with multi-task labels, which achieves superior performance while reducing nearly 2/3 of training costs.

Adaptive Prompt Analysis. We compare the results of different prompt settings to verify the importance of the

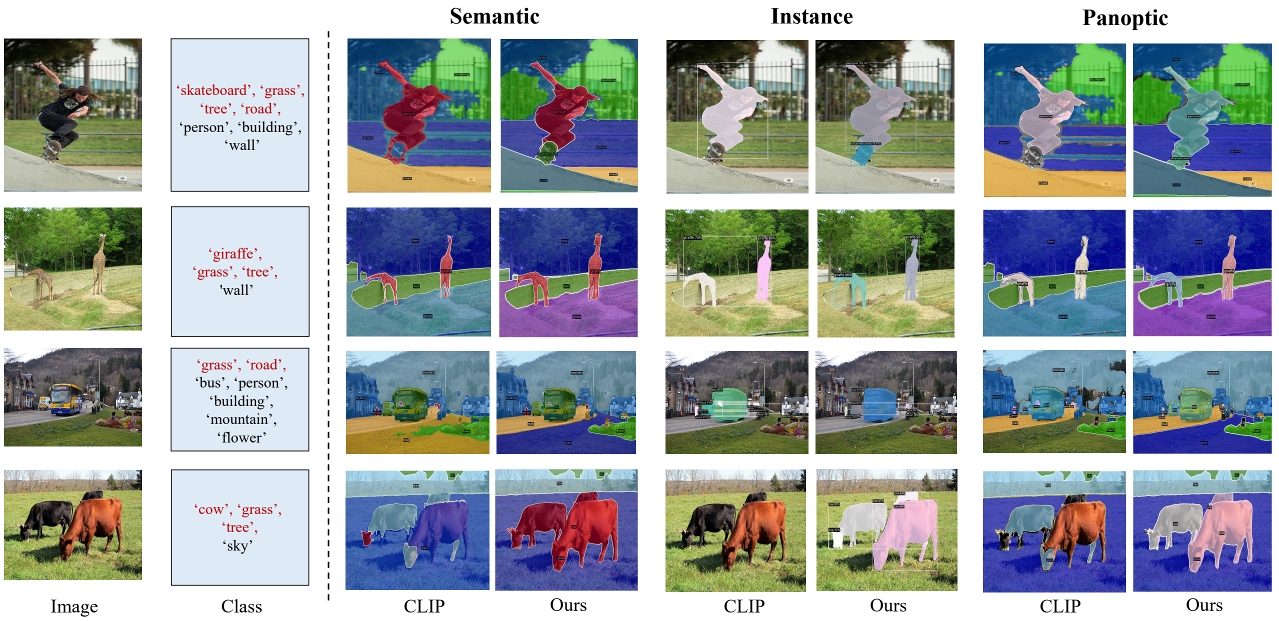

Semantic
Instance
Panoptic
'skateboard', 'grass', 'tree', 'road', 'person', 'building', 'wall'
'giraffe', 'grass', 'tree', 'wall'
'grass', 'road', 'bus', 'person', 'building', 'mountain', 'flower'
'cow', 'grass', 'tree', 'sky'
Image
Class
CLIP
Ours
CLIP
Ours
CLIP
Ours

Figure 3. Qualitative results of the multi-task open vocabulary segmentation. We compare the segmentation results of the proposed FreeSeg and CLIP [26]. The class column represents the class names, where red and black words denote the unseen and seen classes, respectively.

Table 7. Comparison of different training paradigms and prompt solutions on COCO.

Method		**Semantic**		**Instance**		**Panoptic**					
		mIoU^s	**mIoU**^{u}	**mAP**^{s}	**mAP**^{u}	**PQ**^{s}	**SQ**^{s}	**RQ**^{s}	**PQ**^{u}	**SQ**^{u}	**RQ**^{u}
Train Paradigm	Single-Task	41.3	42.9	24.1	12.7	30.1	75.0	37.6	17.5	69.7	21.1
	Multi-Task	41.9	43.3	23.9	14.6	30.4	76.7	38.1	19.2	71.4	24.1
Prompt	Fixed	38.5	33.7	21.4	8.2	26.1	73.0	32.4	12.1	63.6	17.5
	Adaptive	41.9	43.3	23.9	14.6	30.4	76.7	38.1	19.2	71.4	24.1

adaptive prompt for open vocabulary segmentation in Tab.7. The fixed template prompt uses the template sentence "A photo of $\{class\}$." where $\{class\}$ is placed in specific class names. The task name is filled into the template "for $\{task\}$." to get the fixed task prompt. Then the task prompt and the class prompt are encoded into the text features. As shown in Tab.7, the adaptive prompt brings out 3.4% and 9.6% mIoU performance improvement than the fixed prompt regarding seen and unseen classes, respectively. Similarly, the adaptive prompt outperforms the fixed prompt by 0.9% and 2.2% mAP on instance segmentation and by 8.2% and 7.5% PQ on panoptic segmentation. It reveals that adaptive prompt facilitates the prompt to capture task-aware and category-sensitive concepts via learnable parameters.

4.6. Qualitative results

We visualize the qualitative results of the unified open vocabulary segmentation in Fig.3. It can be observed that CLIP fails to segment the instances of some unseen classes like "cow" and "skateboard" in the first and fourth images. However, FreeSeg accurately segments the unseen class regions such as "giraffe" or "grass" for semantic segmentation. These figures show our capability of specifying arbitrary classes in instance and panoptic segmentation. These results demonstrate that FreeSeg is capable of generalizing to arbitrary segmentation categories in universal segmentation tasks.

5. Conclusion

In this paper, we provide a universal framework, *i.e.*, FreeSeg to accomplish unified open-vocabulary segmentation. To the best of our knowledge, we offer the first attempt to employ a single model with the same architecture and inference parameters to accomplish open-vocabulary semantic, instance, and panoptic segmentation. Compared with single-task training, FreeSeg successfully reduced the training cost by about two-thirds and achieved better generalization performance. Only one unified model is needed in real-scene deployment, reducing the inference procedure's computational capacity, memory cost, and bandwidth. We believe our work can provide inspired insight and suggest a new path forward in open-vocabulary segmentation.

References

[1] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2, 5, 6

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 5

[3] 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*, 2017. 1, 2

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 2

[5] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 3, 5

[6] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Perpixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 2, 3

[7] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022. 2, 3, 5, 6

[8] Zheng Ding, Jieke Wang, and Zhuowen Tu. Openvocabulary panoptic segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022. 2, 7

[9] Mark Everingham and John Winn. The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep*, 2007:1–45, 2012. 5

[10] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 540–557. Springer, 2022. 7

[11] Zhangxuan Gu, Siyuan Zhou, Li Niu, Zihan Zhao, and Liqing Zhang. Context-aware feature generation for zeroshot semantic segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1921–1929, 2020. 2, 5, 6

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[13] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7031, 2022. 2

[14] 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 *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[15] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4544–4553, 2020. 5

[16] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 3

[17] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 2

[18] Huixia Li, Chenqian Yan, Shaohui Lin, Xiawu Zheng, Baochang Zhang, Fan Yang, and Rongrong Ji. Pams: Quantized super-resolution via parameterized max scale. In *European Conference on Computer Vision*, pages 564–580. Springer, 2020. 2

[19] Jiashi Li, Qi Qi, Jingyu Wang, Ce Ge, Yujian Li, Zhangzhang Yue, and Haifeng Sun. Oicsr: Out-in-channel sparsity regularization for compact deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7046–7055, 2019. 2

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3

[21] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[22] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 3

[23] Giuseppe Pastore, Fabio Cermelli, Yongqin Xian, Massimiliano Mancini, Zeynep Akata, and Barbara Caputo. A closer look at self-training for zero-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2693–2702, 2021. 2, 5, 6

[24] Jie Qin, Jie Wu, Ming Li, Xuefeng Xiao, Min Zheng, and Xingang Wang. Multi-granularity distillation scheme towards lightweight semi-supervised semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 481–498. Springer, 2022. 2

[25] Jie Qin, Jie Wu, Xuefeng Xiao, Lujun Li, and Xingang Wang. Activation modulation and recalibration scheme for weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2117–2125, 2022. 2

[26] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 5, 6, 7, 8

[27] Shafin Rahman, Salman Khan, and Nick Barnes. Improved visual-semantic alignment for zero-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11932–11939, 2020. 6

[28] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 3

[29] Yuxi Ren, Jie Wu, Xuefeng Xiao, and Jianchao Yang. Online multi-granularity distillation for gan compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6793–6803, 2021. 2

[30] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 1, 2

[31] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020. 5

[32] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A Olshausen, and Trevor Darrell. Fully test-time adaptation by entropy minimization. 2020. 5

[33] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 3

[34] Jie Wu, Tianshui Chen, Hefeng Wu, Zhi Yang, Guangchun Luo, and Liang Lin. Fine-grained image captioning with global-local discriminative objective. *IEEE Transactions on Multimedia*, 23:2413–2427, 2020. 2

[35] Jie Wu, Guanbin Li, Si Liu, and Liang Lin. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12386–12393, 2020. 2

[36] Xin Xia, Jiashi Li, Jie Wu, Xing Wang, Mingkai Wang, Xuefeng Xiao, Min Zheng, and Rui Wang. Trt-vit: Tensorrt-oriented vision transformer. *arXiv preprint arXiv:2205.09579*, 2022. 2

[37] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8256–8265, 2019. 2, 5, 6

[38] 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. 2, 3, 5, 6, 7

[39] Tasweer Ahmad Xuefeng Xiao, Yafeng Yang and Tianhai Chang Lianwen Jin. Design of a very compact cnn classifier for online handwritten chinese character recognition using dropweight and global pooling. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 891–895. IEEE, 2017. 2

[40] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, 2018. 2

[41] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34:10326–10338, 2021. 3

[42] Ye Zheng, Ruoran Huang, Chuanqi Han, Xi Huang, and Li Cui. Background learnable cascade for zero-shot object detection. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 6

[43] Ye Zheng, Jiahong Wu, Yongqiang Qin, Faen Zhang, and Li Cui. Zero-shot instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2593–2602, 2021. 6

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

[45] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3