Exploring Regional Clues in CLIP for Zero-Shot Semantic Segmentation

Yi Zhang[1*] Meng-Hao Guo[2] Miao Wang[1] Shi-Min Hu[2]
[1]State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University
[2]BNRist, Department of Computer Science and Technology, Tsinghua University

2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-5300-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVPR52733.2024.00315

Abstract

CLIP has demonstrated marked progress in visual recognition due to its powerful pre-training on large-scale image-text pairs. However, it still remains a critical challenge: how to transfer image-level knowledge into pixel-level understanding tasks such as semantic segmentation. In this paper, to solve the mentioned challenge, we analyze the gap between the capability of the CLIP model and the requirement of the zero-shot semantic segmentation task. Based on our analysis and observations, we propose a novel method for zero-shot semantic segmentation, dubbed CLIP-RC (CLIP with Regional Clues), bringing two main insights. On the one hand, a region-level bridge is necessary to provide fine-grained semantics. On the other hand, overfitting should be mitigated during the training stage. Benefiting from the above discoveries, CLIP-RC achieves state-of-the-art performance on various zero-shot semantic segmentation benchmarks, including PASCAL VOC, PASCAL Context, and COCO-Stuff 164K. Code will be available at https://github.com/Jittor/JSeg.

1. Introduction

As a foundation task in computer vision, semantic segmentation [4, 8, 11–13, 26, 36, 40, 44, 50, 51] aims to assign each pixel with a semantic class. Limited by technical methods and labeling costs, traditional segmenters can only process scenarios with a limited number of classes. In other words, it can only handle the seen classes in the training set. When it comes to unseen classes, traditional methods seem powerless. However, in practical situations, encountering classes that are not previously seen is unavoidable, which brings challenges to the segmenter. To solve this problem, researchers propose a new research paradigm, called zero-shot semantic segmentation (ZS3) [1, 2, 43], which requires the model trained on seen classes to generalize well to unseen classes. In this paper, we focus on the ZS3 settings.

The rapid development of ZS3 tasks benefits from the progress of the foundation models, especially the CLIP [34] model. The CLIP model is trained on large-scale image-text pairs and shows powerful zero-shot image recognition capability, which is the foundation of ZS3 tasks. In addition to the capability for zero-shot recognition, the ability to localize at the pixel level is also crucial for accomplishing the ZS3 tasks which is an attribute that CLIP lacks. In order to make up for the shortcomings of CLIP in localization, researchers have proposed two streams of approaches: one-stage methods [47, 55] and two-stage methods [7, 14, 33, 46, 52].

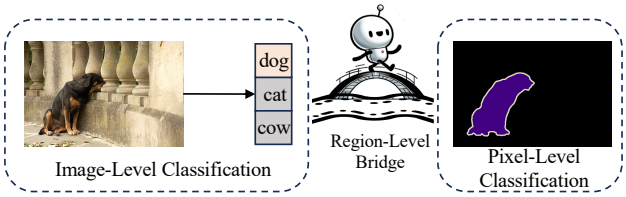

Figure 1. By employing a region-level bridge, the CLIP-RC extends zero-shot capabilities from the image level to the pixel level, thus bridging the gap between image-level recognition and pixel-level semantic segmentation.

Two-stage methods, first, generate initial mask proposals using class-agnostic mask generators. Then, it provides them with semantic information by using the CLIP model. Due to the introduction of an additional class-agnostic segmentation model, these methods have a heavy computation cost. In this paper, we focus on the one-stage methods. One-stage approaches avoid extra computing overhead and finetune the CLIP model for ZS3 directly. There are two critical factors during the fine-tuning process. Firstly, transferring image-level understanding features to pixel-level understanding features are the key to finishing the segmentation task. Secondly, during the fine-tuning process, models will tend to only recognize the classes they see in the tuning process (*a.k.a.,* catastrophic forgetting), which will undermine the model's zero-shot recognition capability. The differences highlighted above are the gaps between the capabilities of the CLIP model and the requirements of ZS3 tasks. Aiming to bridge the above gaps, we present a new method CLIP-RC.

CLIP-RC is motivated by an important observation. As

*Work conducted during an internship at Tsinghua University.

2575-7075/24/$31.00 ©2024 IEEE
DOI 10.1109/CVPR52733.2024.00315

demonstrated in Fig. 2, we simply test the CLIP's capacity on region-level classification. We find CLIP has a strong performance in regional recognition. Region-level recognition capability is a finer-grained recognition capability than the image-level, which is closer to the pixel-level segmentation task. Thus, we believe it can be a suitable bridge to connect CLIP and ZS3 tasks. Building on this, we introduce a new approach for ZS3 named CLIP-RC (CLIP with Regional Clues). This method leverages regional clues, bridging the gap between image-level and pixel-level understanding as illustrated in Fig. 1.

Furthermore, we explored ways to solve the overfitting challenge by adding extra constraints during the tuning process. In detail, we proposed the recovery decoder with recovery loss. It allows CLIP-RC to adapt to ZS3 while minimizing the loss of its inherent generalization abilities. That is, it's designed to strike a delicate balance between task-specific tuning and preserving broad knowledge. This ensures reliable performance across both seen and unseen classes.

Our contributions can be summarized as follows:

- We have introduced a novel framework for ZS3, called CLIP-RC (CLIP with Regional Clues), which can reduce the gap between image-level classification and pixel-level semantic segmentation by introducing a region-level bridge, providing a better solution for ZS3 tasks.
- We introduce a recovery decoder and corresponding recovery loss to mitigate overfitting. This approach effectively balances task-specific knowledge with the model's inherent generalization capabilities.
- By employing CLIP-RC, we've established a new benchmark for state-of-the-art performance in ZS3 across various benchmarks. This achievement surpasses previous methods by large margins.

2. Related Work

Vision-Language Pre-training Model. Vision-language models, like those in [19, 34, 49], are designed to understand the complex connections between visual elements and their textual explanations. A key example of such models is CLIP [34], which uses a large dataset of 400 million internet-sourced image-text pairs to link language with images. CLIP employs contrastive learning to align images and text in a shared feature space. This capability equips CLIP to undertake a variety of computer vision tasks, as indicated in studies like [9, 29, 38, 52], and extends to other areas as well [15, 30, 35, 37, 39]. CLIP's wide-ranging ability to apply its knowledge to various fields and formats highlights its strength, particularly in executing tasks with zero-shot. In this work, we explore how to bridge the gap between the CLIP model pre-trained for image-level classification and the ZS3 task for pixel-level classification. Thus

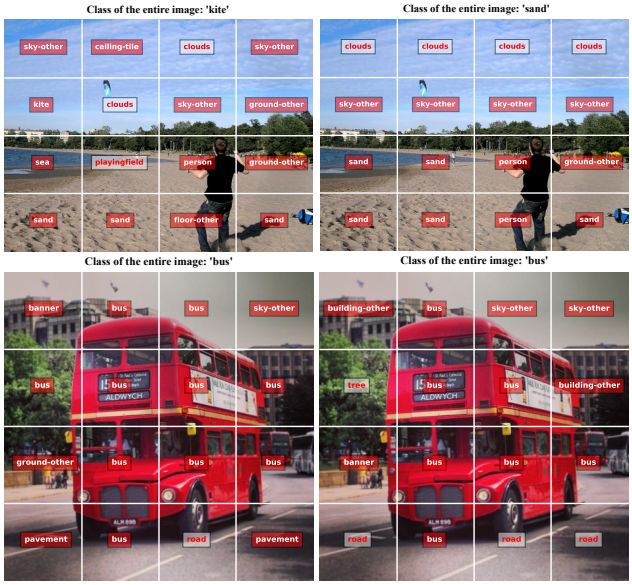

Figure 2. The left side of the figure presents the original CLIP's classification results for a 4×4 grid of regions, demonstrating that CLIP can capture more granular classification information within the image. The right side shows the results of the trained CLIP-RC encoder for classifying the same 4×4 grid of regions. By our design, both seen and unseen classes achieve more accurate classification. Note, the red tags indicate unseen classes.

further unleashing the potential of CLIP in the ZS3.

Model Tuning. Model tuning refines pre-trained models for specific tasks with targeted data, traditionally adjusting all parameters but now focusing on efficiency to avoid overfitting and excessive compute use. The field of Natural Language Processing (NLP) has recently advanced in parameter efficient fine tuning [16, 17, 27], and similar progress is seen in computer vision. Instead of using full fine-tuning as previously done, VPT [20] introduces prompt tuning tokens into transformer layers. At the same time, CoOp [54] transforms the fixed text encoder prompts in the original CLIP model into flexible, trainable vectors. However, CoOp initially tended to overfit the training classes. To solve this, CoCoOp [53] creates unique, condition-based tokens for each image, reducing overfitting. Furthermore, PromptSRC [23] introduces a self-regulation method, further addressing the overfitting issue. CLIP-RC integrates the existing VPT [20] and proposes the use of a region-level bridge to extract region category features, aiming to achieve more suitable tuning for semantic segmentation.

Zero Shot Semantic Segmentation. ZS3 [1, 2, 22, 28, 43, 48] differs from traditional semantic segmentation as it focuses on identifying and segmenting classes that are not labeled or seen during training. The rise of pre-trained vision-language models has greatly impacted this field, making completing zero-shot tasks much easier. Models like ZSSeg [46] and ZegFormer [7] have proposed a two-stage approach for ZS3. The first stage extracts mask pro-

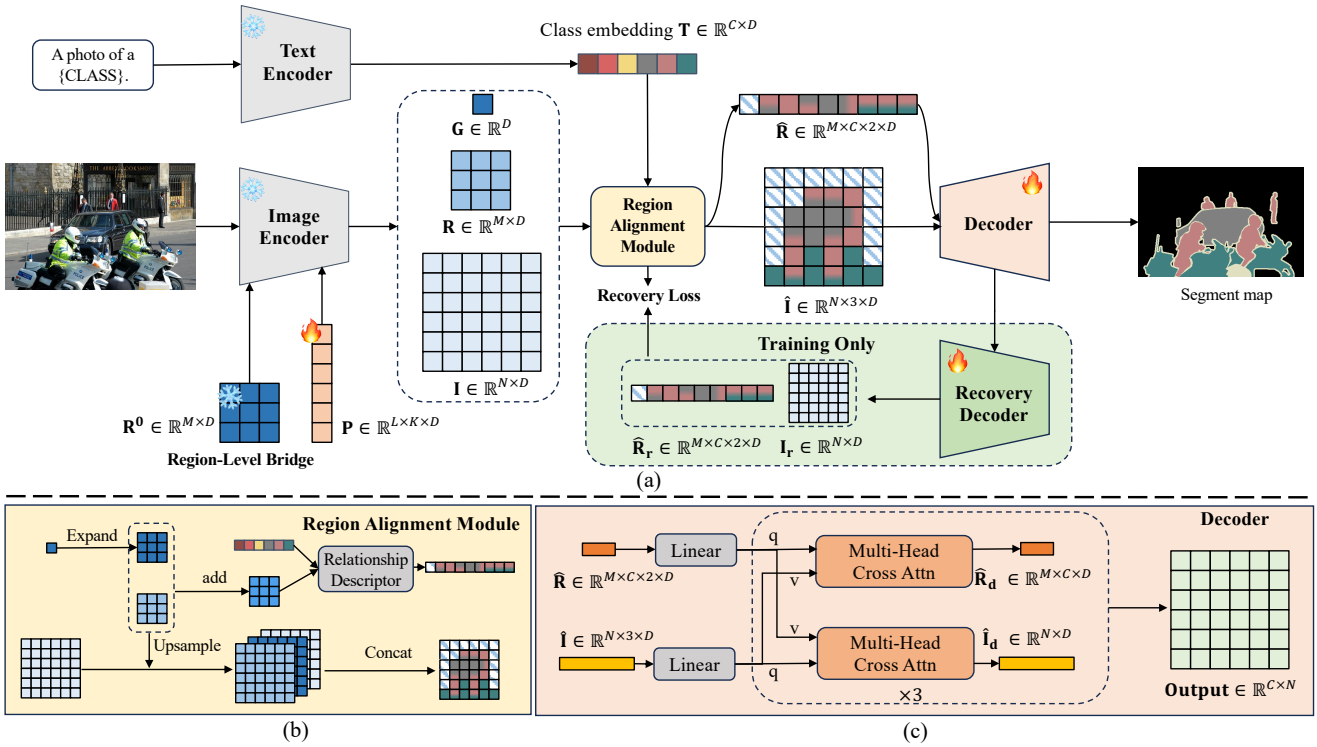

Figure 3. The framework of our proposed CLIP-RC. (a) The image is input into the image encoder, which then yields the image feature $\mathbf{I}$, the region category information feature $\mathbf{R}$ extracted via region-level bridging, and the global feature $\mathbf{G}$. Subsequently, these three feature sets are aligned and combined with the text embedding $\mathbf{T}$ to generate the regional relationship descriptors. Finally, a decoder for semantic segmentation then utilizes these features to infer and generate a segmentation map. To prevent overfitting during training, the recovery decoder and recovery loss are employed. (b) The region alignment module. (c) Detail of the decoder architecture.

posals with a class-agnostic mask generator, and the second stage employs the CLIP model for zero-shot classification of the masked image sections. MaskCLIP+ [52] improves upon previous ZS3 methods significantly by incorporating pseudo-labeling and self-training. However, these two-stage methods can be computationally expensive. To avoid this, ZegCLIP [55] innovates further by combining visual prompt tuning with relationship descriptors for a one-stage ZS3 inference process. In addition, the task of open vocabulary semantic segmentation, similar to the ZS3 task, also has been explored through various methods [3, 21, 24, 25, 41, 42, 45, 47]. It is worth mentioning that CLIPSelf [42] has a similar motivation as ours. The difference is that it transfers regional features to the student model through knowledge distillation. In this paper, we explore the ZS3 approach from two distinct perspectives, aiming to identify and bridge the existing gap between current single-stage methods and ZS3.

3. Method

3.1. Method Overview

As illustrated in Fig. 3(a), the CLIP-RC has three key components: the Region-Level Bridge (RLB), Region Alignment Module (RAM), and Recovery Decoder with Recovery Loss (RDL). First, as detailed in Sec. 3.2, the image is fed into the encoder to obtain global features, image features, and region category features extracted by RLB. In Sec. 3.3, we use RAM to align these feature sets and create image features with finer-grained categorical features. The features are also merged with the text embedding to obtain regional relationship descriptors. Following this, a decoder for semantic segmentation uses these features to predict and create a segmentation map. To minimize overfitting during training, the RDL, discussed in Sec. 3.4, is used to ensure a balance between learning task-specific features and general knowledge.

3.2. Region-Level Bridge

In our approach, the Region-Level Bridge (RLB) is a key element to connect image-level and pixel-level representations. Specifically, as illustrated in Fig. 4. RLB captures the regional category features of the image and facilitates classification at a regional granularity.

Formally, we construct the input for the first ViT layer of CLIP's visual encoder as shown in Eq. (1). Each element represents distinct aspects of the input data,

$$\mathbf{X}^0 = [\mathbf{G}^0, \mathbf{P}^0, \mathbf{I}^0, \mathbf{R}^0], \tag{1}$$

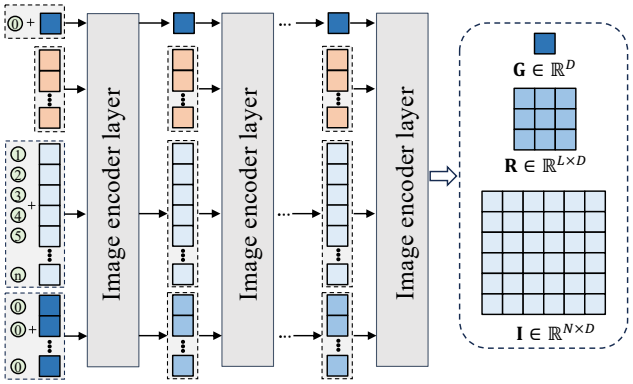

Figure 4. This illustrates the RLB within the image encoder layers. The input consists of a [CLS] token, deep prompt tokens, image features, and RLB (from top to bottom). Each token in RLB is responsible for a distinct image region, thereby yielding finer-grained categorical features about image regions. Upon completion of the inference process, the image features are obtained, along with the [CLS] token that contains global image features and regional category features.

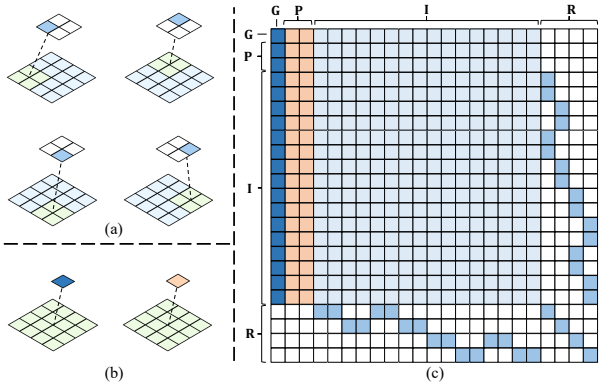

Figure 5. Visualization of attention mask in CLIP. (a) Each token in the RLB is responsible for a 2×2 region within the 4×4 image features. (b) For [CLS] token and deep prompt tuning tokens, each token interacts with all image features, including interactions among themselves. (c) Visualization of the attention mask during self-attention, where the white blocks indicate masking and no interaction.

where $\mathbf{G}^0 \in \mathbb{R}^{1\times D}$ represents the [CLS] token for the visual encoder, which is designed to capture the global image feature, where D represents the dimension of the features. The $\mathbf{P}^0 \in \mathbb{R}^{K\times D}$ represents the K deep prompt tuning tokens [20] for the first layer, selected from the deep prompt tuning token set $\mathbf{P} \in \mathbb{R}^{L\times K\times D}$ that are task-specific learnable parameters into the input space. $\mathbf{I}^0 \in \mathbb{R}^{N\times D}$ denotes the image features with added positional encoding. Lastly, $\mathbf{R}^0 \in \mathbb{R}^{M\times D}$ symbolizes the initial RLB with M tokens. Each token in $\mathbf{R}^0$ adopts the weights of $\mathbf{G}^0$ as its initial weights, serving as a bridge between image-level and pixel-level features.

Following the construction of the input, we proceed with the extraction of image features. For a sequence of image features with a length of N, the original size is $\sqrt{N} \times \sqrt{N} \times D$. We utilize mask attention to accomplish the extraction of regional category features of image regions through the RLB. As illustrated in Fig. 5, the attention mask $\mathbf{Mask} \in \mathbb{R}^{E\times E}$ defines the computational direction of the RLB, where $E = 1 + K + N + M$. Each token in the RLB is responsible for extracting features from $\frac{\sqrt{N}}{\sqrt{M}} \times \frac{\sqrt{N}}{\sqrt{M}}$ patches. In this way, each token in RLB is responsible for one region in the original whole feature, which is an intermediate granularity between image-level and pixel-level. This feature extraction method is almost the same as the feature extraction when CLIP does classification on images.

After obtaining the attention mask, the masked attention operations within the visual encoder can be formulated as:

$$\begin{aligned}\mathbf{X}^{l+1} &= \mathcal{V}^{l+1}_{\mathrm{MHSA}}(\mathbf{X}^l) \\ &= \mathrm{softmax}(\frac{QK^T}{\sqrt{d_v}} + \mathbf{Mask})V,\end{aligned} \tag{2}$$

where, $\mathbf{X}^l$ and $\mathbf{X}^{l+1}$ are the input and output of the transformer $l+1$ layer $\mathcal{V}^{l+1}_{\mathrm{MHSA}}$ respectively. It is important to note that to ensure the generalizability of the RLB for unseen classes, its gradient is not updated.

The final outputs of the visual encoders, denoted as $\mathbf{X}^L$, are given by:

$$\mathbf{X}^L = [\mathbf{G}, \_, \mathbf{I}, \mathbf{R}], \tag{3}$$

where, $\mathbf{G}$ signifies the global category feature of an image by the [CLS] token, $\mathbf{I}$ denotes the features of the image extracted by the visual encoder and $\mathbf{R}$ contains the region category features by the RLB.

3.3. Region Alignment Module

Building upon the outputs of the CLIP visual encoder, we introduce the Region Alignment Module (RAM), as illustrated in Fig. 3(b). This module is specifically designed to further align different feature sets. By aligning the $\mathbf{G}$ that gathers the global image context with the $\mathbf{R}$ that are specific to the region and image features $\mathbf{I}$, the multi-scale spatial features and fine-grained category features present in the input data can be fully utilized.

To align the feature sets, we reshape $\mathbf{R}$ back to a dimension of $\sqrt{M} \times \sqrt{M} \times D$, representing the category features of regions in the image. We then upsample both $\mathbf{G}$, containing global image category features, and $\mathbf{R}$ to the size of image features, and concatenate them with $\mathbf{I}$:

$$\hat{\mathbf{I}} = \mathrm{Concat}(\mathrm{Upsample}(\mathbf{G}), \mathrm{Upsample}(\mathbf{R}), \mathbf{I}). \tag{4}$$

Further, to generate text embeddings with generalized capabilities for unseen classes, we employ the Relationship Descriptor [55]. This involves integrating the priors of the RLB and the global [CLS] token into the text embeddings, resulting in multiple, robust text embeddings with different regional priors, *i.e.* regional relationship descriptors. We

first fuse the RLB with the global [CLS] token:

$$\mathbf{R_a} = (\text{Upsample}(\mathbf{G}) + \mathbf{R}), \tag{5}$$

$\mathbf{R_a}$ is shaped as $M \times D$, and then:

$$\hat{\mathbf{R}} = \text{Concat}[\mathbf{R_t}, \mathbf{T}] = \text{Concat}[\mathbf{T} \odot \mathbf{R_a}, \mathbf{T}], \tag{6}$$

$\mathbf{T} \in \mathbb{R}^{C \times D}$ represents the original text embeddings extracted via CLIP's text encoder, and C is the number of classes. The $\hat{\mathbf{R}} \in \mathbb{R}^{M \times C \times 2 \times D}$ are the regional relationship descriptors.

3.4. Recovery Decoder With Recovery Loss

To mitigate overfitting in zero-shot learning with a CLIP-based model, we introduce the training-only Recovery Decoder with Recovery Loss (RDL). This RDL focuses on balancing task-specific adaptation with the retention of general knowledge by adding extra constraints during the tuning process

Initially, we deploy a decoder for semantic segmentation, as shown in Fig. 3(c). The aligned image features $\hat{\mathbf{I}}$ and region-specific text queries $\hat{\mathbf{R}}$, derived from Sec. 3.3, pass through a linear layer, aligning them in dimension D. They are then processed as follows:

$$\hat{\mathbf{I}}_\mathbf{d}, \hat{\mathbf{R}}_\mathbf{d} = \mathcal{D}_{\text{MHCA}}(\hat{\mathbf{I}}, \hat{\mathbf{R}}), \mathcal{D}'_{\text{MHCA}}(\hat{\mathbf{R}}, \hat{\mathbf{I}}), \tag{7}$$

where, $\mathcal{D}_{\text{MHCA}}$ and $\mathcal{D}'_{\text{MHCA}}$ denotes the decoder for semantic segmentation with multi-head cross attention, and $\hat{\mathbf{I}}_\mathbf{d} \in \mathbb{R}^{N \times D}$ and $\hat{\mathbf{R}}_\mathbf{d} \in \mathbb{R}^{M \times C \times D}$ are the image features and region-specific text queries respectively, used for segmentation. The segmentation map $\mathbf{Output} \in \mathbb{R}^{C \times N}$ is obtained by averaging the outputs:

$$\mathbf{Output_i} = \hat{\mathbf{R}}_\mathbf{d} \hat{\mathbf{I}}_\mathbf{d}^T \quad \text{for all } i \in \{1, \ldots, M\}. \tag{8}$$

Then, during training, a recovery decoder recovers the features extracted by the decoder into features with strong generalization. The network architecture of the recovery decoder is completely identical to that of the semantic segmentation decoder. They are recovered as follows:

$$\hat{\mathbf{I}}_\mathbf{r}, \hat{\mathbf{R}}_\mathbf{r} = \mathcal{RD}_{\text{MHCA}}(\hat{\mathbf{I}}_\mathbf{d}, \hat{\mathbf{R}}_\mathbf{d}), \mathcal{RD}'_{\text{MHCA}}(\hat{\mathbf{R}}_\mathbf{d}, \hat{\mathbf{I}}_\mathbf{d}), \tag{9}$$

$\hat{\mathbf{I}}_\mathbf{r} \in \mathbb{R}^{N \times D}$ and $\hat{\mathbf{R}}_\mathbf{r} \in \mathbb{R}^{M \times C \times D}$ represent features after recovery of $\mathbf{I}$ and $\hat{\mathbf{R}}$ respectively. To ensure that the outputs of the recovery decoder composed of $\mathcal{RD}_{\text{MHCA}}$ and $\mathcal{RD}'_{\text{MHCA}}$ align well with the original features from the backbone network, we propose a recovery loss for use with the recovery decoder. This recovery loss is designed to help the decoder for semantic segmentation strike a balance between learning the specifics of the task at hand and maintaining a broad base of general knowledge, thereby alleviating the problem of overfitting to unseen classes. The equation for this loss is:

$$\mathcal{L}_{\text{recovery}} = \sum_{i=1}^{n} |\hat{\mathbf{I}}_{\mathbf{r}i} - \mathbf{I}_i| + \sum_{i=1}^{n} |\hat{\mathbf{R}}_{\mathbf{r}i} - \hat{\mathbf{R}}_i|. \tag{10}$$

3.5. Loss Function

To further reduce overfitting during our training, we use a method called Non-mutually Exclusive Loss (NEL) [55]. This method combines Sigmoid activation with Binary Cross Entropy (BCE) loss, allowing for the independent prediction of probabilities for different classes.

Additionally, we incorporate a recovery loss as discussed in Section Sec. 3.4. The total loss our model aims to minimize is a combination of these two types of losses, represented by the equation:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{NEL}} + \beta \cdot \mathcal{L}_{\text{recovery}}, \tag{11}$$

where, α and β are weights that balance the contributions of NEL and recovery loss, respectively.

4. Experiments

4.1. Dataset

PASCAL VOC 2012 Dataset provides an augmented training set of 10,582 images, alongside a validation set consisting of 1,449 images. In our work, we exclude the background class and categorize the 20 available classes into 15 seen classes and 5 unseen classes.

COCO-Stuff164K Dataset covers 80 thing classes, 91 stuff classes, and a single class designated for unlabeled elements. It comprises a training subset featuring 118,287 images, alongside a validation subset consisting of 5,000 images. The entire dataset is further divided into 156 seen classes and 15 unseen classes.

PASCAL Context Dataset contains 59 foreground classes and a "background" class. The training set and validation set contain 4,996 and 5,104 images, respectively. The dataset is divided into 50 seen classes (including "background") and 10 unseen classes.

4.2. Implementation Details

Our experiments were conducted using the Jittor [18] and PyTorch [32] frameworks, with code based on the MMSegmentation library [6]. We used the ViT-B/16 model from CLIP, training on 8 NVIDIA RTX 3090 GPUs. The batch size was set at 16 for all datasets, using an input image size of 512×512. In the inductive setting, we trained on the Pascal VOC 2012, Pascal Context, and COCO-Stuff 164K datasets for 40K, 40K, and 80K iterations, respectively. For the transductive setting, we loaded the weight of the checkpoint from the middle of the inductive training for each dataset, then trained each from scratch for 20K, 20K, and 40K iterations, respectively.

4.3. Evaluation Protocol

Continuing from the established methodology for ZS3 [43, 52, 55], all classes C of a dataset are divided into a group

Methods	PASCAL VOC 2012				PASCAL Context				COCO-Stuff 164K			
	pAcc	hIoU	mIoU(S)	mIoU(U)	pAcc	hIoU	mIoU(S)	mIoU(U)	pAcc	hIoU	mIoU(S)	mIoU(U)
SPNet [43]	-	26.1	78.0	15.6	-	-	-	-	-	14.0	35.2	8.7
ZS3 [2]	-	28.7	77.3	17.7	52.8	15.8	20.8	12.7	-	15.0	34.7	9.5
CaGNet [10]	80.7	39.7	78.4	26.6	-	21.2	24.1	18.5	65.6	18.2	33.5	12.2
SIGN [5]	-	41.7	75.4	28.9	-	-	-	-	-	20.9	32.3	15.5
Joint [1]	-	45.9	77.7	32.5	-	20.5	33.0	14.9	-	-	-	-
ZegFormer [7]	-	73.3	86.4	63.6	-	-	-	-	-	34.8	36.6	33.2
zsseg [46]	90.0	77.5	83.5	72.5	-	-	-	-	60.3	37.8	39.3	36.3
DeOP [14]	-	80.8	88.2	74.6	-	-	-	-	-	38.2	38.0	38.4
ZegCLIP [55]	94.6	84.3	91.9	77.8	**76.2**	49.9	46.0	54.6	62.0	40.8	40.2	41.4
CLIP-RC (Ours)	**95.8**	**88.4**	**92.8**	**84.4**	**76.2**	**51.9**	**47.5**	**57.3**	**63.1**	**41.2**	**40.9**	**41.6**

Table 1. Comparison with the SOTA methods in the **inductive setting**.

Methods	PASCAL VOC 2012				PASCAL Context				COCO-Stuff 164K			
	pAcc	hIoU	mIoU(S)	mIoU(U)	pAcc	hIoU	mIoU(S)	mIoU(U)	pAcc	hIoU	mIoU(S)	mIoU(U)
SPNet+ST [43]	-	38.8	77.8	25.8	-	-	-	-	-	30.3	34.6	26.9
ZS5 [2]	-	33.3	78.0	21.2	49.5	23.4	27.0	20.7	-	16.2	34.9	10.6
CaGNet+ST [10]	81.6	43.7	78.6	30.3	-	-	-	-	56.8	19.5	35.6	13.4
STRICT [31]	-	49.8	82.7	35.6	-	-	-	-	-	34.8	35.3	30.3
zsseg+ST [46]	88.7	79.3	79.2	78.1	-	-	-	-	63.8	41.5	39.6	43.6
MaskCLIP+ [52]	-	87.4	88.8	86.1	-	53.3	44.4	**66.7**	-	45.0	38.1	54.7
FreeSeg [33]	-	86.9	82.6	91.8	-	-	-	-	-	45.3	**42.2**	49.1
ZegCLIP+ST* [55]	96.2	91.1	92.3	89.9	**77.4**	54.0	47.2	63.2	69.2	48.5	40.7	59.9
CLIP-RC(Ours)	**97.0**	**93.0**	**93.9**	**92.2**	77.2	**55.1**	**48.1**	64.5	**69.9**	**49.7**	42.0	**60.8**

Table 2. Comparison with the SOTA methods in the **transductive setting**. ST represents Self-Training. * denotes the results obtained from our retraining of the method on the Pascal Context dataset.

of seen classes C_S and a group of unseen classes C_U, with $C_S \cap C_U = \emptyset$. During training, only the seen classes (C_S) have labels. Furthermore, in the inductive setting of ZS3, the model is trained without any knowledge of the unseen classes C_U, including their labels, and names. This setting closely mirrors practical inference scenarios, where the model may be tested on classes not seen during the training. Contrastingly, in the transductive setting of ZS3, the names of unseen classes are known before testing. This setting can enhance the performance of these unseen classes and reduce the dependence on data annotation for practical scenarios.

For the evaluation metric, we evaluate our model's performance using standard metrics in segmentation: Mean Intersection-over-Union (mIoU) and pixel-wise classification accuracy (pAcc). mIoU is separately reported for seen (mIoU(S)) and unseen (mIoU(U)) classes. Additionally, the Harmonic Mean IoU (hIoU) ensures a balanced evaluation of the model's performance on both seen and unseen classes, computed using the formula:

$$\text{hIoU} = \frac{2 \times \text{mIoU(S)} \times \text{mIoU(U)}}{\text{mIoU(S)} + \text{mIoU(U)}} \tag{12}$$

4.4. Comparison with the State-of-the-art

Our proposed method, CLIP-RC, has been extensively evaluated on various benchmarks, displaying outstanding performance in both inductive and transductive settings. Its ability to effectively transfer segmentation capabilities to unseen classes proves its effectiveness for ZS3 tasks.

In the inductive setting, detailed in Tab. 1, CLIP-RC outperforms the current SOTA model, ZegCLIP [55], by a significant margin. This is particularly true in handling unseen classes in the PASCAL VOC 2012 dataset, where our hIoU reaches 88.4%, a notable 4.1% improvement. This enhancement is most apparent in the recognition of unseen classes, underlining the improved segmentation capability of our method. Similar superiority is observed in the PASCAL Context and COCO-Stuff 164K datasets. We also showcase visual results in Fig. 6 from the COCO-Stuff 164K dataset, where CLIP-RC accurately distinguishes between various unseen classes, such as 'playing field' and 'cloud', and 'cardboard', among others.

In the transductive setting, as detailed in Tab. 2, CLIP-RC uses self-training to achieve groundbreaking results. It attained an hIoU of 93.0% on the PASCAL VOC 2012 dataset, surpassing the previous best model by 1.9%. This strong performance is also evident in the PASCAL Context and COCO-Stuff 164K datasets. Notably, in the transductive setting, CLIP-RC outperforms the SOTA trained in fully supervised environments through self-training. This highlights the effectiveness of our method in accurately identifying seen classes and successfully generalizing them to unseen classes.

To demonstrate the upper limit of our model's capabil-

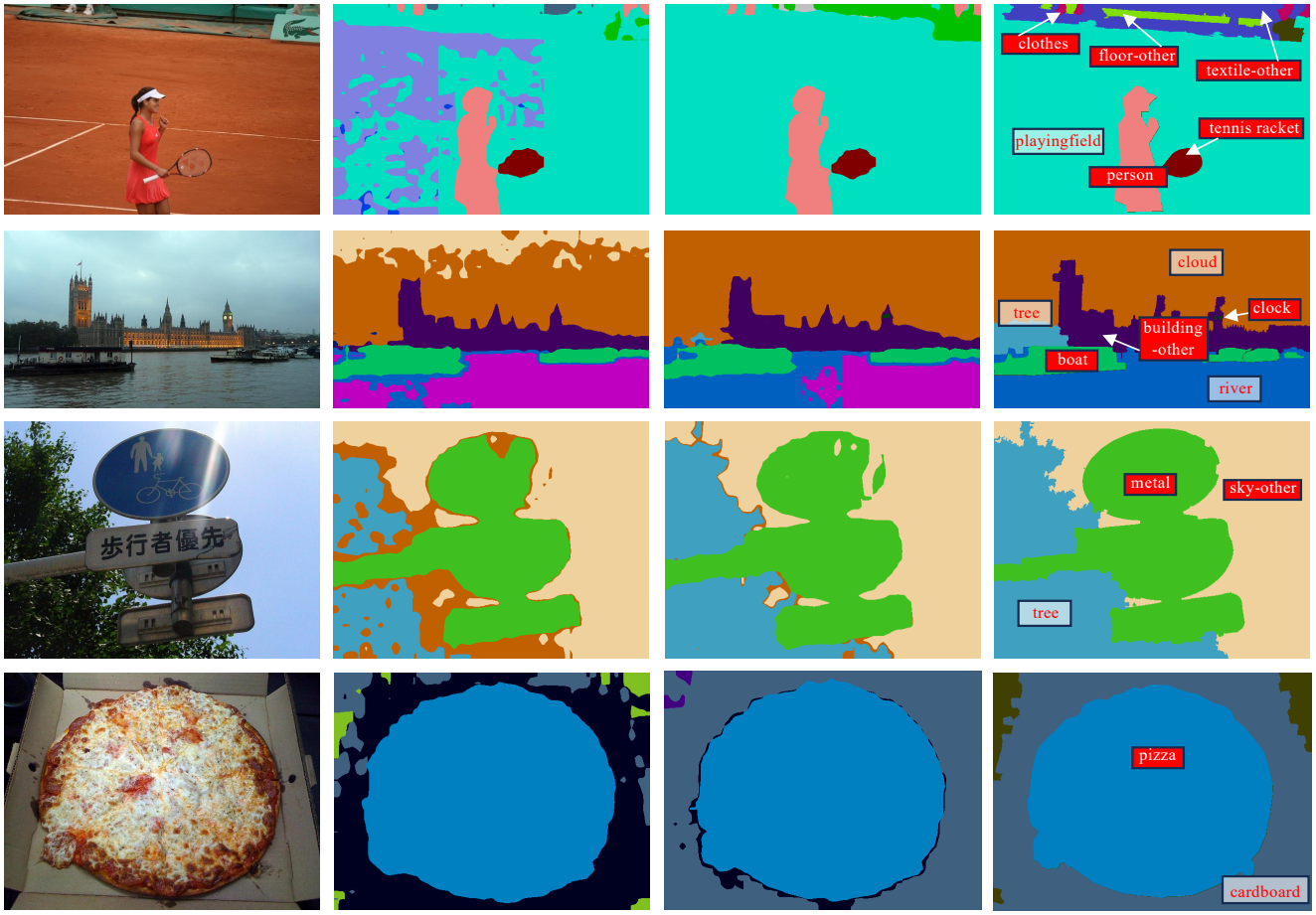

Figure 6. Visualization results on the COCO-Stuff 164K dataset. Columns from left to right: (1) the test images, (2) results using the current SOTA method [55], (3) results using CLIP-RC (Ours), and (4) ground truth. The red tags indicate unseen classes.

Methods	PASCAL VOC 2012				PASCAL Context				COCO-Stuff 164K			
	pAcc	hIoU	mIoU(S)	mIoU(U)	pAcc	hIoU	mIoU(S)	mIoU(U)	pAcc	hIoU	mIoU(S)	mIoU(U)
ZegCLIP*	96.3	91.6	92.4	90.9	**77.4**	56.2	46.9	70.0	69.9	49.6	40.7	63.2
CLIP-RC (Ours)	**97.1**	**93.7**	**94.1**	**93.4**	76.9	**56.7**	**47.1**	**71.3**	**70.8**	**51.4**	**42.9**	**64.1**

Table 3. Comparison with the SOTA methods in the **fully supervised setting**, * denotes the results obtained from our retraining of the method on the Pascal Context dataset

Method	Pascal Context			Pascal Context59		
	pAcc	mIoU	mAcc	pAcc	mIoU	mAcc
Zegformer [7]	42.3	29.3	56.6	-	-	-
ZegCLIP [55]	60.9	41.2	68.4	68.4	47.5	69.7
CLIP-RC (ours)	**62.1**	**42.3**	**68.8**	**70.9**	**49.2**	**69.9**

Table 4. Cross-dataset generalization efficacy from COCO-Stuff 164K to PASCAL Context.

ities, Tab. 3 presents our fully supervised training results. The CLIP-RC consistently exhibits a higher ceiling compared to the former SOTA, substantiating its versatility beyond the zero-shot setting.

Moreover, we delved into the cross-dataset generalization capabilities of our model. In additional experiments presented in Tab. 4, we train on the seen classes from COCO-Stuff 164K and test on PASCAL Context and Context59 (without 'background'), illustrating our approach's adaptability and cross-dataset generalization ability.

4.5. Ablation Study

In our ablation study on the PASCAL VOC 2012 dataset within an inductive setting, we thoroughly assessed the unique contributions of each component in our proposed design. Initially, we conducted a broad evaluation of each component's effectiveness, followed by detailed ablation tests for each specific module.

RLB	RAM	RDL	PASCAL VOC 2012			
			pAcc	hIoU	mIoU(S)	mIoU(U)
✗	✗	✗	94.6	84.3	91.9	77.8
✓	✗	✗	95.2	85.4	91.5	80.0
✓	✓	✗	95.4	87.4	92.5	82.9
✓	✓	✓	**95.8**	**88.4**	**92.8**	**84.4**

Table 5. Ablation Studies of Key Components: RLB signifies Region-Level Bridge, RAM denotes Region Alignment Module, and RDL represents the Recovery Decoder with Recovery Loss.

Tokens Number	Update	GFLOPs	PASCAL VOC 2012			
			pAcc	hIoU	mIoU(S)	mIoU(U)
16	✓	126.7	95.8	87.9	93.3	83.1
1	✗	119.8	95.3	86.9	92.5	82.0
4	✗	121.2	95.6	87.3	92.6	82.6
16	✗	126.7	95.8	88.4	92.8	84.4
64	✗	148.9	**96.0**	89.1	**93.0**	85.6
256	✗	238.2	**96.0**	**89.2**	92.7	**86.0**
1024	✗	609.2	OOM			

Table 6. Ablation studies of the region-level bridge

[CLS] token	RLB	PASCAL VOC 2012			
		pAcc	hIoU	mIoU(S)	mIoU(U)
✗	✗	94.9	85.6	91.5	80.5
✓	✗	95.4	86.6	92.1	82.0
✗	✓	95.7	87.0	92.9	81.8
✓	✓	**95.8**	**88.4**	**92.8**	**84.4**

Table 7. Ablation studies of region alignment module

Loss Type	Number of Layer	PASCAL VOC 2012			
		pAcc	hIoU	mIoU(S)	mIoU(U)
L1 Loss	1	**95.8**	**88.4**	**92.8**	**84.4**
L2 Loss	1	95.2	86.7	91.8	81.9
KD Loss	1	95.4	86.4	92.2	81.2
L1 Loss	0	95.0	86.6	91.3	82.1
L1 Loss	2	95.4	87.1	92.4	82.4
L1 Loss	3	95.3	87.0	91.6	82.8

Table 8. Ablation studies of the recovery decoder and recovery loss

In Tab. 5, we benchmarked our model against the previous SOTA [55], to measure the impact of each module on our model's overall effectiveness. The results demonstrate that each component significantly enhances our model's performance in ZS3 tasks.

Region-Level Bridge (RLB). Our ablation study in Tab. 6 examined whether updating the RLB improves performance and what the ideal number of these tokens is. We found that updating the RLB can lead to overfitting on seen classes, negatively affecting the performance of unseen classes. The study also investigated how changing the number of tokens affects the RLB. Generally, increasing the number of tokens enhances the model's performance up to a certain point. Beyond this computational limit, issues like out-of-memory (OOM) errors can occur, as we noted with 1024 tokens. To balance effectiveness and computational efficiency, we decided to use 16 tokens.

Region Alignment Module (RAM). Our investigation into the RAM scrutinizes which features should be preserved during fusion as specified in Eq. (4). Results in Tab. 7 indicate that aligning and fusing both the [CLS] token and the RLB with the image's original features is most effective. The combined use of these tokens yields the best results, optimizing the performance of our model.

Recovery Decoder with Recovery Loss (RDL). In our exploration of the RD, we evaluated the impact of different recovery loss types and the number of decoder layers on model performance in Tab. 8. Among various loss functions, L1 loss emerged as the most effective. As for the number of layers in the RD, we found that increasing layers does not necessarily improve results; a balance between task-specific knowledge and general knowledge must be reached. Additionally, we considered removing the Recovery decoder entirely, computing the L1 loss directly between features used for segmentation and the original features. However, this direct application impacted the model's fitting ability and led to decreased performance.

5. Conclusion

In this work, we presented CLIP-RC, a novel one-stage method for ZS3. Our approach successfully bridges the gap between image-level classification and pixel-level segmentation by introducing regional clues. This method demonstrates the potential of leveraging finer-grained recognition capabilities by CLIP for ZS3 tasks. By integrating a recovery decoder and recovery loss, we addressed the issue of overfitting, striking a balance between maintaining the inherent generalization abilities of the CLIP model and adapting it for the ZS3 task. Our experimental results have shown that CLIP-RC not only performs robustly on seen classes but also exhibits remarkable performance on unseen classes. This dual capability is crucial for real-world applications where encountering novel classes is common. We hope this work paves the way for more efficient and effective solutions in the rapidly growing fields of ZS3.

Limitations. CLIP-RC can refine the granularity of regions, as shown in Tab. 6. Finer granularity can improve IoU to some extent, but it also increases computational load. Furthermore, since this work is based on CLIP, the performance of the method also depends on the effectiveness of the pre-training of CLIP in visual-language alignment.

Acknowledgement. This work was supported by the National Key Research and Development Program of China (project No. 2021ZD0112902), the National Natural Science Foundation of China (project Nos. 62220106003, 62372025), the Research Grant of Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology and the Fundamental Research Funds for the Central Universities.

References

[1] Donghyeon Baek, Youngmin Oh, and Bumsub Ham. Exploiting a joint embedding space for generalized zero-shot semantic segmentation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9516–9525. IEEE, 2021. 1, 2, 6

[2] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 466–477, 2019. 1, 2, 6

[3] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11165–11174, 2023. 3

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, pages 833–851. Springer, 2018. 1

[5] Jiaxin Cheng, Soumyaroop Nandi, Prem Natarajan, and Wael Abd-Almageed. SIGN: spatial-information incorporated generative network for generalized zero-shot semantic segmentation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9536–9546. IEEE, 2021. 6

[6] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 5

[7] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11573–11582. IEEE, 2022. 1, 2, 6, 7

[8] Lixue Gong, Yiqun Zhang, Yunke Zhang, Yin Yang, and Weiwei Xu. Erroneous pixel prediction for semantic image segmentation. *Computational Visual Media*, 8(1):165–175, 2022. 1

[9] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 2

[10] Zhangxuan Gu, Siyuan Zhou, Li Niu, Zihan Zhao, and Liqing Zhang. Context-aware feature generation for zero-shot semantic segmentation. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 1921–1929. ACM, 2020. 6

[11] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 1

[12] Meng-Hao Guo, Zheng-Ning Liu, Tai-Jiang Mu, and Shi-Min Hu. Beyond self-attention: External attention using two linear layers for visual tasks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(5):5436–5447, 2023.

[13] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. *Computational Visual Media*, 9(4):733–752, 2023. 1

[14] Cong Han, Yujie Zhong, Dengjie Li, Kai Han, and Lin Ma. Open-vocabulary semantic segmentation with decoupled one-pass network. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 1086–1096. IEEE, 2023. 1, 6

[15] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: zero-shot text-driven generation and animation of 3d avatars. *ACM Trans. Graph.*, 41(4):161:1–161:19, 2022. 2

[16] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 2790–2799. PMLR, 2019. 2

[17] 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. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 2

[18] Shi-Min Hu, Dun Liang, Guo-Ye Yang, Guo-Wei Yang, and Wen-Yang Zhou. Jittor: a novel deep learning framework with meta-operators and unified graph execution. *Sci. China Inf. Sci.*, 63(12), 2020. 5

[19] 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 *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 4904–4916. PMLR, 2021. 2

[20] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIII*, pages 709–727. Springer, 2022. 2, 4

[21] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for zero-shot open-vocabulary segmentation. *CoRR*, abs/2306.09316, 2023. 3

[22] Naoki Kato, Toshihiko Yamasaki, and Kiyoharu Aizawa. Zero-shot semantic segmentation via variational mapping. In *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea*

(South), October 27-28, 2019, pages 1363–1370. IEEE, 2019. 2

[23] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 15144–15154. IEEE, 2023. 2

[24] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 7633–7642. IEEE, 2023. 3

[25] 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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 7061–7070. IEEE, 2023. 3

[26] Zheng Lin, Zhao Zhang, Ziyue Zhu, Deng-Ping Fan, and Xialei Liu. Sequential interactive image segmentation. *Computational Visual Media*, 9(4):753–765, 2023. 1

[27] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9):195:1–195:35, 2023. 2

[28] Xianglong Liu, Shihao Bai, Shan An, Shuo Wang, Wei Liu, Xiaowei Zhao, and Yuqing Ma. A meaningful learning method for zero-shot semantic segmentation. *Sci. China Inf. Sci.*, 66(11), 2023. 2

[29] Timo Lüddecke and Alexander S. Ecker. Image segmentation using text and image prompts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7076–7086. IEEE, 2022. 2

[30] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 13482–13492. IEEE, 2022. 2

[31] 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 *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, pages 2693–2702. Computer Vision Foundation / IEEE, 2021. 6

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas K"opf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035, 2019. 5

[33] Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Yuxi Ren, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, and Xingang Wang. Freeseg: Unified, universal and open-vocabulary image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19446–19455. IEEE, 2023. 1, 6

[34] 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 *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 1, 2

[35] Aditya Sanghi, Hang Chu, Joseph G. Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18582–18592. IEEE, 2022. 2

[36] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651, 2017. 1

[37] Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3d gaussians for open-vocabulary scene understanding. *arXiv preprint arXiv:2311.18482*, 2023. 2

[38] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5198–5215. Association for Computational Linguistics, 2022. 2

[39] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 3825–3834. IEEE, 2022. 2

[40] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. PVT v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 1

[41] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, et al. Towards open vocabulary learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[42] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *CoRR*, abs/2310.01403, 2023. 3

[43] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero- and few-label semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8256–8265. Computer Vision Foundation / IEEE, 2019. 1, 2, 5, 6

[44] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jos'e M. 'Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 12077–12090, 2021. 1

[45] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18113–18123. IEEE, 2022. 3

[46] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXIX*, pages 736–753. Springer, 2022. 1, 2, 6

[47] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. SAN: side adapter network for open-vocabulary semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(12):15546–15561, 2023. 1, 3

[48] Hui Zhang and Henghui Ding. Prototypical matching and open set rejection for zero-shot semantic segmentation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6954–6963. IEEE, 2021. 2

[49] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *CoRR*, abs/2304.00685, 2023. 2

[50] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6230–6239. IEEE Computer Society, 2017. 1

[51] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 6881–6890. Computer Vision Foundation / IEEE, 2021. 1

[52] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII*, pages 696–712. Springer, 2022. 1, 2, 3, 5, 6

[53] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16795–16804. IEEE, 2022. 2

[54] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348, 2022. 2

[55] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting CLIP for zero-shot semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 11175–11185. IEEE, 2023. 1, 3, 4, 5, 6, 7, 8