GROUNDED CONTRASTIVE LEARNING FOR OPEN-WORLD SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

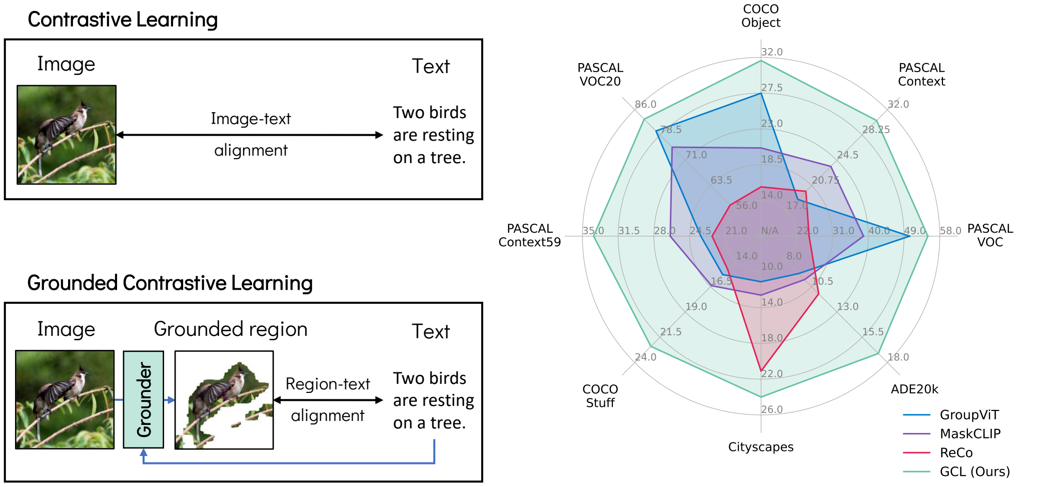
Contrastive learning (CL) with large-scale image-text paired data has made great strides in open-world image recognition. The progress raises attraction to open-world semantic segmentation—aiming at learning to segment arbitrary visual concepts in images. Existing open-world segmentation methods adopt CL to learn diverse visual concepts and adapt its image-level understanding to the segmentation task. However, while CL-based existing methods have shown impressive results, conventional CL is limited in considering image-text level alignment without explicit optimization of region-text level alignment, thus leading to a sub-optimal solution for the segmentation task. In this paper, we propose a novel *Grounded Contrastive Learning (GCL)* framework to directly align a text and regions described by the text. Our method generates a segmentation mask associated with a given text, extracts grounded image embedding from the masked image region, and aligns it with text embedding via GCL. The framework encourages a model to directly improve the quality of generated segmentation masks. In addition, for a rigorous and fair comparison, we present a unified evaluation protocol with widely used 8 semantic segmentation datasets. GCL achieves state-of-the-art zero-shot segmentation performance with large margins in all datasets. The code will be released publicly available.

1 INTRODUCTION

Semantic segmentation aims to identify the class of every pixel in an image. This task has been studied extensively over decades and has shown impressive improvements (Chen et al., 2017; Strudel et al., 2021; Xie et al., 2021). However, the existing methods are inherently limited in segmenting a small number of object categories presented in the training dataset. Such limitation mainly comes from the difficulty in collecting costly pixel-level annotations for diverse categories.

The recent emergence of large-scale image-text paired data (Changpinyo et al., 2021; Schuhmann et al., 2021; Byeon et al., 2022) from the web paves the way for open-world recognition which aims to learn a capability of identifying arbitrary semantic categories in the open world. Since the texts contain a global semantic description for the paired images, the large-scale image-text paired data can provide knowledge for diverse semantic categories. For instance, CLIP (Radford et al., 2021) learns visual and language representations using contrastive learning with the large-scale image-text data, and shows its open-world image classification capability to identify arbitrary image class labels by a simple image-text matching strategy.

Inspired by such success of open-world image classification, a few works (Zhou et al., 2022; Shin et al., 2022; Xu et al., 2022; Liu et al., 2022) for the open-world semantic segmentation have been suggested based on image-text contrastive learning. However, the absence of pixel-level dense annotations in the image-text data still makes the open-world segmentation task challenging. Thus, the existing methods mainly rely on the transferability of the image-text alignment capability learned during the training to perform region-text alignment at inference time. More specifically, MaskCLIP (Zhou et al., 2022) leverages a pre-trained model such as CLIP, simply modifies its last layer to extract more precise patch-level visual features, and finally generates segmentation masks by patch-text matching. GroupViT (Xu et al., 2022) and ViL-Seg (Liu et al., 2022) propose to cluster region-level visual features into distinct groups and perform matching between clustered regions and text to generate segmentation mask. While the existing methods have shown impressive results even



(a) Concept comparison.

(b) Performance comparison in 8 datasets.

Figure 1: (a) A conceptual comparison between conventional contrastive learning (CL) and grounded contrastive learning (GCL). CL-based methods learn image-level alignment and implicitly learn a grounding capability, which facilitates generating segmentation masks. In contrast, in GCL, the grounder is incorporated to learn region-text alignment from image-text data, which is trained in an end-to-end manner. (b) As a result, the proposed method significantly outperforms existing methods in all 8 segmentation benchmark datasets.

from the training with image-text alignment, they are still limited in that no explicit optimization for region-text alignment is considered as depicted in Figure 1a.

To tackle such limitation, we propose **G**rounded **C**ontrastive **L**earning (GCL) framework, which directly learns to align text and region of interest from the image-text data. Our key idea is to incorporate a text grounding capability within contrastive learning as shown in Figure 1a. To be specific, given a pair of image and text, we first obtain a segmentation mask indicating text-grounded regions using a grounder, compute grounded region embeddings using the segmentation mask and finally perform contrastive learning between the text and the grounded region. By making the contrastive loss be affected from the quality of segmentation, GCL enables to train the grounder and directly improves the quality of text-region level alignment. In addition, we present a unified evaluation protocol with widely used 8 semantic segmentation datasets and compare existing methods—evaluated with their own protocol—in the same setting using the protocol. GCL achieves state-of-the-art zero-shot segmentation performance with large margins in all datasets as presented in Figure 1b.

The main contribution of this paper is summarized as follows:

- We introduce a novel framework for the open-world semantic segmentation, named Grounded Contrastive Learning, which directly optimizes alignment between text and region of interest, thus learning to generate more precise segmentation masks even from image-text paired data.
- We present a unified evaluation protocol and re-evaluate recent open-world segmentation models for a fair and direct comparison.
- We achieve the new state-of-the-art zero-shot segmentation performance on 8 segmentation datasets with large margins compared to existing methods.

2 Related Works

Semantic Segmentation In recent years, we have witnessed the dramatic advancement and the success of semantic segmentation in deep learning, including FCN (Long et al., 2015), U-Net (Ronneberger et al., 2015), DeepLab (Chen et al., 2017), and the very recent transformer-based methods (Strudel et al., 2021; Xie et al., 2021). However, these conventional segmentation models are constrained to fixed-label sets defined by the labels specified in the training dataset, for example, Cityscapes (Cordts et al., 2016), PASCAL VOC (Everingham et al., 2010), Context59 (Mottaghi et al., 2014) of 19/20/59 classes, and they are incompetent to segment any categories beyond the predefined ones. In natural images, segmentation targets vary more than predefined categories (typically under 100) in terms of categories, attributes (*e.g.*, colors, materials, etc.), geometric locations (*e.g.*, "car on the left"), and relation to the environment ("baby is sitting on the bench"), etc. The traditional semantic segmentation methods are incapable of dealing with these problems. To alleviate discrepancy between training and test categories, zero-shot segmentation methods (Bucher et al., 2019; Li et al., 2020) have been proposed that handle unseen categories at test time. As another line of research, weakly supervised semantic segmentation methods (Ahn & Kwak, 2018; Lee et al., 2019; Chang et al., 2020) are proposed to train the model using image-level category supervision instead of pixel-level supervision. However, these segmentation models are still suffering from addressing many concepts. In contrast, our method tries to endow the segmentation model with open-world recognition capability, which can segment any category by learning a wide range of concepts via large-scale image-text pairs (Sharma et al., 2018; Changpinyo et al., 2021; Schuhmann et al., 2021; Byeon et al., 2022).

Open-world Semantic Segmentation Open-world models have recently gained popularity since traditional fully supervised models cannot handle classes that are not defined during testing. The pioneering work Contrastive Language-Image Pre-training (CLIP, Radford et al. 2021) ushered in the era of open-world image recognition using large-scale image-text pairs. It learns the alignment between an image and a text at training time, then transfers it to the zero-shot classification by aligning the image and given set of classes at inference time. The advent of CLIP enables open-world setting in various fields such as image captioning (Hessel et al., 2021), object detection (Gu et al., 2022; Zhong et al., 2022), and semantic segmentation.

However, while contrastive learning (CL) allows zero-shot image classification by learning image-text level alignment, semantic segmentation needs additional consideration since segmentation tasks require capturing specific regions of the image corresponding to the target class. In order to transfer the pre-trained model with image-text alignment into the segmentation task, the model needs to learn region-text level alignment (*i.e.*, capturing which region in the image is aligned with the text).

Existing open-world semantic segmentation studies have taken a strategy to overlook this issue. Instead of learning region-level alignment directly in pre-training stage, they transfer image-level alignment to region-level by heuristic modification (Zhou et al., 2022; Shin et al., 2022) or clustering (Xu et al., 2022; Liu et al., 2022). MaskCLIP (Zhou et al., 2022) proposes to obtain a dense image embedding from CLIP image encoder through heuristic modification of the last attention layer. Even though it has several limitations, such as low output resolution or noisy segmentation results, they show it is a simple yet effective way to obtain an initial segmentation map for refinement. ReCo (Shin et al., 2022) proposes an advanced refinement method based on MaskCLIP, by retrieval and co-segmentation. Clustering-based methods (Xu et al., 2022; Liu et al., 2022) learn representations using CL with image-text pairs. They compute region-level image embedding by clustering sub-region embeddings. These approaches also have shown impressive results but have several limitations: 1) the learning objective is still image-level alignment, 2) clustering sub-region image embeddings is independent of query text, and has compelling difficulties in handling arbitrary concepts. In summary, existing methods indirectly use CL as a training objective to address region-level alignment problems. To tackle this problem, we propose a novel region-level alignment objective named Grounded Contrastive Learning (GCL), which provides an objective for region-level alignment. By solving the problem explicitly, we achieve state-of-the-art performance with large margins on all evaluation benchmarks.

3 METHODS

3.1 FROM CONTRASTIVE LEARNING TO GROUNDED CONTRASTIVE LEARNING

Open-world semantic segmentation is a task that aims to learn a model having zero-shot segmentation capability for arbitrary visual concepts. Our main goal is to develop an open-world segmentation algorithm using image-text paired data only. This objective is hard to achieve particularly because there are no explicit learning signals (*i.e.*, pixel-level dense annotations) for the segmentation of texts (or objects included in the texts). Therefore, for random image $\mathbf{x}^V$ and text $\mathbf{x}^T$, existing methods typically learn models parametrized by θ to maximize the mutual information between paired

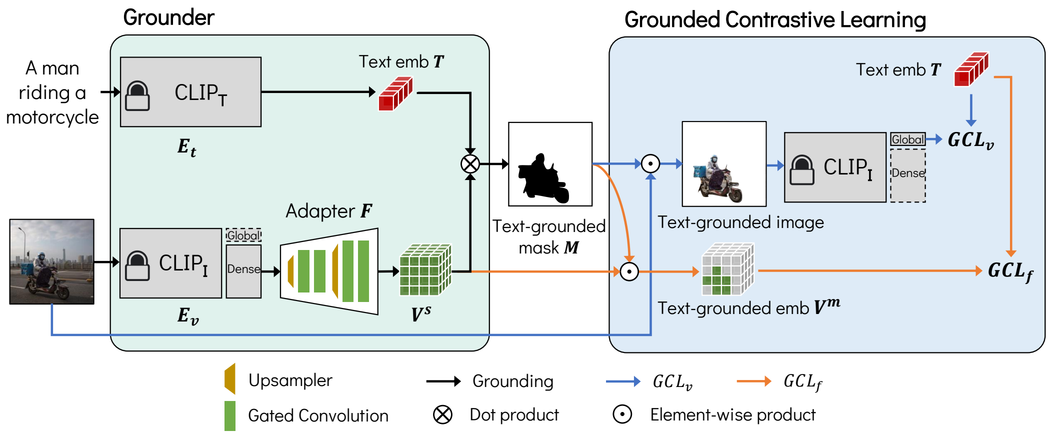


Figure 2: **Overall training pipeline of GCL.** CLIP$_\text{T}$ and CLIP$_\text{I}$ indicate CLIP text and image encoders, respectively. The proposed GCL framework provides an objective to learn region-text level alignment. CLIP encoders are frozen and we train the adapter network only. After the training, GCL block is discarded, and only Grounder block is used to generate the text-grounded segmentation mask.

images and texts (Oord et al., 2018; Radford et al., 2021) as follows:

$$\arg\max_\theta I_\theta(\mathbf{x}^V; \mathbf{x}^T). \tag{1}$$

This objective encourages the model to learn the alignment between texts and images, however, at inference time, the learned model requires generating segmentation masks for texts of arbitrary visual concepts by computing region-text alignments. Such alignment-level discrepancy between training and inference time may lead the model to sub-optimal solution. With this in consideration, to bridge the gap between the objective of conventional contrastive learning and requirement in the segmentation process, we propose Grounded Contrastive Learning (GCL) which incorporates a text grounding process to directly align text and regions of interest; for the text grounding we introduce a grounder that is in charge of generating segmentation masks for the given texts. In a nutshell, GCL learns a model to maximize mutual information between texts and text-grounded regions as follows:

$$\arg\max_\theta I_\theta(\mathbf{m} \cdot \mathbf{x}^V; \mathbf{x}^T), \tag{2}$$

where $\mathbf{m}$ is a text-grounded mask of random variable indicating the regions described by a given text. Compared to the contrastive learning (CL) that implicitly learns a grounding capability, GCL has a clear advantage of explicit learning grounding capability from the end-to-end trainable grounder.

In the rest of this section, we first explain the overall pipeline of GCL including model architecture of the grounder and the mask generation process. Then, we describe how we define losses using the generated mask to train our open-world grounder.

3.2 Overall Pipeline

Figure 2 illustrates our overall training pipeline. For an input batch of paired texts $\mathbf{X}^T$ and images $\mathbf{X}^V$, GCL first performs a grounding process to identify text-grounded regions for a text via a grounder. The grounder consists of two encoders: 1) image encoder is in charge of providing a single (L2-normalized) global feature as well as dense patch-level features and 2) text encoder provides a (L2-normalized) text embedding feature. In practice, as shown in Figure 2, we adopt CLIP (Radford et al., 2021) to initialize two encoders. To preserve and exploit the CLIP's useful knowledge learned during large-scale pre-training, we freeze the pre-trained CLIP and introduce an adapter network for dense patch-level features. To be specific, given features computed using CLIP text encoder E_t, CLIP image encoder E_v, and the adapter network F, segmentation masks are computed via position-wise dot product between text embedding and dense patch-level features as follows:

$$\mathbf{T} = E_t\left(\mathbf{X}^T\right) \quad \text{and} \quad \mathbf{V}^g, \mathbf{V}^d = E_v\left(\mathbf{X}^V\right), \tag{3}$$

$$\mathbf{V}^s = F(\mathbf{V}^d), \tag{4}$$

$$\mathbf{M}_{i,j} = \sigma\left(w \cdot \mathbf{t}_j^\top \mathbf{V}_i^s + b\right), \tag{5}$$

```
# x_t[B, L]        - minibatch of texts
# x_v[B, 3, H, W] - minibatch of images
# text_encoder    - return text emb (t)
# image_encoder   - return global image emb and dense feature (v_g, v_d)
# adapter         - convert v_d into dense image emb (v_s)
# proj            - scalar projection and sigmoid layer

t = text_encoder(x_t)  # t[B, C]
_, v_d = image_encoder(x_v)  # v_d[B, L, C]
v_s = adapter(v_d)  # v_s[B, C, H, W]
mask = proj(einsum("ichw,jc->ijhw", v_s, t))  # [B, B, H, W]

pos_mask = mask[arange(B), arange(B)].unsqueeze(1)  # [B, 1, H, W]
pos_mask_b = gumbel_sigmoid(pos_mask)  # binarized positive mask
v_g, _ = image_encoder(pos_mask_b * x_v)  # v_g[B, C]
s = v_g @ t.T
gcl_v = info_nce(s)  # InfoNCE loss

w = mask / mask.sum((2, 3), keepdim=True)
grounded_embedding = einsum("ichw,ijhw->ijc", v_s, w)
s = einsum("ijc,ic->ij", grounded_embedding, t)
gcl_f = info_nce(s)  # InfoNCE loss

gcl_loss = gcl_v + gcl_f
```

Figure 3: PyTorch-like pseudo code for the core implementation of GCL.

where $\mathbf{T} \in \mathbb{R}^{B\times C}$, $\mathbf{V}^g \in \mathbb{R}^{B\times C}$ and $V^d \in \mathbb{R}^{B\times L\times C}$ are a normalized text embedding, a normalized global image embedding and dense image features from CLIP, $\mathbf{V}^s \in \mathbb{R}^{B\times C\times H\times W}$ is a normalized dense image embedding by the adapter network, and $\mathbf{M} \in \mathbb{R}^{B\times B\times H\times W}$ is segmentation masks between images and texts in the batch. B, C and L mean a batch size, the embedding dimension size and the number of patches, respectively. σ is a sigmoid function, and w, b are learned scalar projection parameters.

Then, the segmentation masks are used to extract grounded-image features. Replacing the global image embedding in conventional contrastive learning with grounded-image embedding makes our model become to learn text-region level alignment in end-to-end manner. In the following sections, we describe how the generated mask $\mathbf{M}$ is used in the GCL framework.

3.3 Training Objectives

Recall that the main idea of GCL is to define positive pairs for contrastive learning as grounded regions and texts. For this purpose, we define GCL losses in two different levels—image-level and feature-level—using the generated masks $\mathbf{M}$ for all pairs of images and texts in a batch; the detailed pseudo code to compute GCL losses is provided in Figure 3. In addition, we employ area prior loss and smooth prior loss to further improve the quality of generated masks.

Image-level GCL loss. Given a pair of image $\mathbf{X}^V_i$ and text $\mathbf{X}^T_i$, one intuitive way to obtain a feature for grounded regions is directly inferring a masked image given by a generated segmentation mask. To make the whole process end-to-end trainable, we first compute a binarized mask $\mathbf{M}^b_{i,i}$ from the generated mask $\mathbf{M}_{i,i}$ using Gumbel-Softmax (Jang et al., 2017) technique and obtain a differentiable masked image via element-wise multiplication of the given image and the binarized mask. Then, we feed the masked image into the image encoder, $\tilde{\mathbf{v}}^g_i, \tilde{\mathbf{v}}^d_i = E_v\left(\mathbf{M}^b_{i,i} \cdot \mathbf{X}^V_i\right)$, and use the extracted global vector $\tilde{\mathbf{v}}^g_i$ as the text-grounded image embedding. We compute a cosine similarity matrix between text-grounded image embeddings and text embeddings in a batch by $S^m_{i,j} = \tilde{\mathbf{v}}^{g\top}_i \mathbf{t}_j$. Finally, based on the similarity matrix, we define the image-level GCL loss $\mathcal{L}_{\text{GCL}_v}$ to make the representations of positive image-text pairs similar to each other while the representations of negative pairs dissimilar using the symmetric version of InfoNCE (Oord et al., 2018; Radford et al., 2021):

$$\mathcal{L}_{\text{GCL}_v} = \text{InfoNCE}\left(\mathbf{S}^m\right), \tag{6}$$

$$\text{InfoNCE}\left(\mathbf{S}\right) = -\frac{1}{2}\left(\frac{1}{B}\sum_i^B \log\frac{\exp(S_{i,i}/\tau)}{\sum_j^B \exp(S_{i,j}/\tau)} + \frac{1}{B}\sum_i^B \log\frac{\exp(S_{i,i}/\tau)}{\sum_j^B \exp(S_{j,i}/\tau)}\right), \tag{7}$$

where τ is a learnable temperature.

Feature-level GCL loss. While the image-level GCL loss drives a model to generate segmentation masks for the associated texts (*i.e.*, text of positive pairs in the batch), we observe that it is not enough to suppress a segmentation mask for the region that is not described in text, especially for the salient region. This raises the need to optimize negative masks obtained from the unrelated text (*i.e.*, text of negative pairs in the batch). However, it is infeasible to compute the image-level GCL loss for negative masks due to the high computational cost of inferring grounded image embeddings. Motivated by Girshick (2015), we introduce the feature-level GCL loss to effectively compute features of the negative masks. Specifically, for dense image embeddings $\mathbf{V}_i^s \in \mathbb{R}^{C \times H \times W}$ and a text embedding vector $\mathbf{t}_j$, we compute feature-level grounded image embedding $\mathbf{v}_{i,j}^f \in \mathbb{R}^C$ by

$$\mathbf{v}_{i,j}^f = \frac{\sum_{h,w} M_{i,j,h,w} \cdot \mathbf{v}_{i,:,h,w}^s}{\sum_{h,w} M_{i,j,h,w}}. \tag{8}$$

Then, we compute a cosine similarity $S_{i,j}^f = {\mathbf{v}_{i,j}^f}^\top \mathbf{t}_j$ between all pairs of text embeddings and feature-level text-grounded image embeddings in the batch. Finally, we define the feature-level GCL loss to suppress the grounding of negative texts using the similarity matrix as follows:

$$\mathcal{L}_{\text{GCL}_f} = \text{InfoNCE}\left(\mathbf{S}^f\right). \tag{9}$$

Area prior loss. We observe that a model trained with the two above losses sometimes converges to a trivial solution—generating a segmentation mask where the whole region is activated—as CL. To address the issue, we leverage priors for segmentation area into our GCL framework. To be specific, for the positive masks (masks from positive pairs) $\mathbf{M}^+$ and the negative masks (masks from negative pair) $\mathbf{M}^-$, we denote the area of positive and negative masks by $\overline{\mathbf{M}^+}$ and $\overline{\mathbf{M}^-}$, respectively. The area prior loss is defined by setting a prior to each area:

$$\mathcal{L}_{\text{area}} = \left\| p^+ - \mathbb{E}\left[\overline{\mathbf{M}^+}\right] \right\|_1 + \left\| p^- - \mathbb{E}\left[\overline{\mathbf{M}^-}\right] \right\|_1, \tag{10}$$

where p^+ and p^- are positive and negative area priors. For the area prior of negative masks, intuitively, we can expect the area of the negative masks to be 0.0 (=p^-). Choosing the positive area prior is a bit more tricky, since it may vary depending on the training dataset. We measure the average positive area in CC3M dataset (Sharma et al., 2018) using previous open-world segmentation model, MaskCLIP (Zhou et al., 2022). As a result, we set p^+ to 0.4.

Smooth prior loss. In the image-text dataset, a text generally describes the salient object or semantic in the paired image. Based on the observation, we assume that most of the text-described region is smooth rather than noisy (like including holes). We employ total variation (TV) regularization loss (Rudin et al., 1992) for the smooth prior. We apply total variation loss to both mask and dense image embedding:

$$\mathcal{L}_{\text{tv}} = \|\mathbf{M}\|_{\text{TV}} + \|\mathbf{V}^s\|_{\text{TV}}, \tag{11}$$

where $\|\cdot\|_{\text{TV}}$ is the anisotropic TV norm.

Final loss. Our final loss function is defined by:

$$\mathcal{L} = \lambda_{\text{GCL}} \mathcal{L}_{\text{GCL}} + \lambda_{\text{area}} \mathcal{L}_{\text{area}} + \lambda_{\text{tv}} \mathcal{L}_{\text{tv}}, \tag{12}$$

where $\mathcal{L}_{\text{GCL}} = \mathcal{L}_{\text{GCL}_v} + \mathcal{L}_{\text{GCL}_f}$, and λ_{GCL}, λ_{area}, λ_{tv} are hyperparameters to balance three losses.

3.4 Refinement

Another important component for semantic segmentation under limited supervision is the refinement process. Some previous works propose their own refinement methods (Zhou et al., 2022; Shin et al., 2022), but we simply adopt existing method, pixel-adaptive mask refinement (PAMR, Araslanov & Roth (2020)). We do not use dense CRF (Krähenbühl & Koltun, 2011) due to its heavy computational burden, despite its widely known effectiveness.

4 EXPERIMENTS

4.1 EXPERIMENT SETTINGS

Unified evaluation protocol. In the open-world semantic segmentation field, there is no standard evaluation protocol yet. Thus, existing methods conduct evaluation using their own protocols such as different data processing strategies on different datasets; surprisingly, even for the same dataset, the target classes are sometimes different across methods. Inspired by this, we present a unified evaluation protocol for a fair and direct comparison of algorithms. We design the protocol following open-world scenario, where prior access to the target data before evaluation is not allowed. Under this scenario, the proposed protocol prohibits dataset-specific hyperparameters (HPs) or tricks, *e.g.*, class name expansion or rephrase, leading to performance overestimation. For example, we observe that the if target class of "person" can be expanded (or rephrased) to its sub-concepts (*e.g.*, man, woman, worker, rider, etc.) depending on dataset, GCL can get significant performance gains. With this in consideration, we evaluate models using unified class names from default version of MMSegmentation (Contributors, 2020) without class name expansion or rephrase. All other evaluation settings follow Xu et al. (2022), where the input image is resized to have a shorter side of 448. Additionally, dense CRF (Krähenbühl & Koltun, 2011) is not used identically due to its expensive computational cost. We employ mean intersection-over-union (mIoU) as a performance metric, which is a standard metric in semantic segmentation field.

Benchmark datasets and comparison methods. To provide extensive evaluation, we first collect all previously used benchmark datasets in the existing methods and add widely used ADE20K dataset for segmentation, resulting in total 8 benchmark datasets—categorized into 2 groups: (i) w/ background class: PASCAL VOC (21 classes), PASCAL Context (60 classes), COCO-Object (80 thing and 1 background classes, Xu et al. (2022)) and (ii) w/o background class: PASCAL VOC20 (20 classes, Everingham et al. (2010)), PASCAL Context59 (59 classes, Mottaghi et al. (2014)), COCO-Stuff (171 classes, Caesar et al. (2018)), Cityscapes (19 evaluation classes, Cordts et al. (2016)), and ADE20k (150 classes, Zhou et al. (2019)). Note that open-world segmentation methods perform segmentation based on class name. Thus, the background class, which does not sufficiently describe the semantic itself, needs additional considerations, *e.g.*, probability thresholding instead of use of "background" text itself. The datasets with background class evaluate this aspect. We compare GCL with all existing open-sourced methods: GroupViT (Xu et al., 2022), MaskCLIP (Zhou et al., 2022), and ReCo (Shin et al., 2022) using the unified protocol. Additional comparison with non-open sourced methods is given in Appendix.

Implementation details. For the grounder, we use CLIP ViT-B/16 model where the size of input images is 224×224 and the patch size is 16×16. Following MaskCLIP (Zhou et al., 2022), we modify the last attention layer of CLIP image encoder to acquire the dense embedding representing local semantics. The adapter network consists of four gated convolution blocks with two upsampling interpolation, where the output of convolution is gated by learned gating parameter and added to the skip connection; the process of gated convolution can be written as: $\mathbf{x}' = \mathbf{x} + \tanh(g) \cdot \text{Conv}(\mathbf{x})$ where $\mathbf{x}$ is input. The detailed architecture of the adapter is provided in Appendix. We use CC3M (Sharma et al., 2018) and CC12M datasets (Changpinyo et al., 2021) for training. The loss weights of $\lambda_{\text{GCL}} = 0.1, \lambda_{\text{area}} = 0.4, \lambda_{\text{tv}} = 1.0$ are used. We train a model with a batch size of 1024 and learning rate of 7.5×10^{-5} for total $50,000$ iterations with $15,000$ warmup steps and cosine schedule. AdamW optimizer (Loshchilov & Hutter, 2018) is used with a weight decay of 0.05.

4.2 MAIN EXPERIMENTS

In Table 1, we first compare the existing methods based on the proposed unified protocol. We compare two checkpoints of GroupViT and two methods of MaskCLIP, due to the performance varying according to the dataset. Between the existing methods, GroupViT (Xu et al., 2022) achieves the best on average, but the best method varies according to each dataset. Interestingly, GroupViT shows the best performance on the object-oriented datasets (VOC, VOC20, and COCO-Object), but the performance tends to decrease as the stuff classes are dominated in the target dataset. For example, GroupViT (RedCaps) presents the worst or second worst performance in Context, Cityscapes, and ADE20K. In contrast, MaskCLIP shows the best performance in stuff-oriented datasets such

Table 1: **Zero-shot segmentation performance comparison on 8 semantic segmentation datasets.** mIoU metric is used in every experiment. We highlight **the best** and second-best results. Object, Stuff, City, ADE indicates COCO-Object, COCO-Stuff, Cityscapes, ADE20K datasets, respectively. MaskCLIP† indicates their baseline method without additional refinement techniques (key smoothing and prompt denoising). The YFCC and RedCaps of GroupViT indicate their training dataset in addition to CC12M.

	with background class			without background class					
Methods	VOC	Context	Object	VOC20	Context59	Stuff	City	ADE	Avg.
GroupViT (YFCC)	49.5	19.0	24.3	74.1	20.8	12.6	6.9	8.7	27.0
GroupViT (RedCaps)	50.4	18.7	27.5	79.7	23.4	15.3	11.1	9.2	29.4
MaskCLIP†	29.3	21.1	15.5	53.7	23.3	14.7	21.6	10.8	23.7
MaskCLIP	38.8	23.6	20.6	74.9	26.4	16.4	12.6	9.8	27.9
ReCo	25.1	19.9	15.7	57.7	22.3	14.8	21.1	11.2	23.5
GCL (Ours)	**55.0**	**30.4**	**31.6**	**83.2**	**33.9**	**22.4**	**24.0**	**17.1**	**37.2**
	(+4.6)	**(+6.8)**	**(+4.1)**	**(+3.5)**	**(+7.5)**	**(+6.0)**	**(+2.4)**	**(+5.9)**	**(+7.8)**

as Context, Context59, COCO-Stuff. We conjecture it is benefited from the leveraging a large-scale pre-trained CLIP. Furthermore, their refinement techniques, key smoothing and prompt denoising, improve the average performance $+4.2$ mIoU, but significantly degrade the performance in Cityscapes dataset ($21.6 \rightarrow 12.6$). It may suggest a limitation of the heuristic refinement and the need for data-driven approach. The signficant performance degradation of MaskCLIP between VOC20 and VOC may imply the need of consideration for background class.

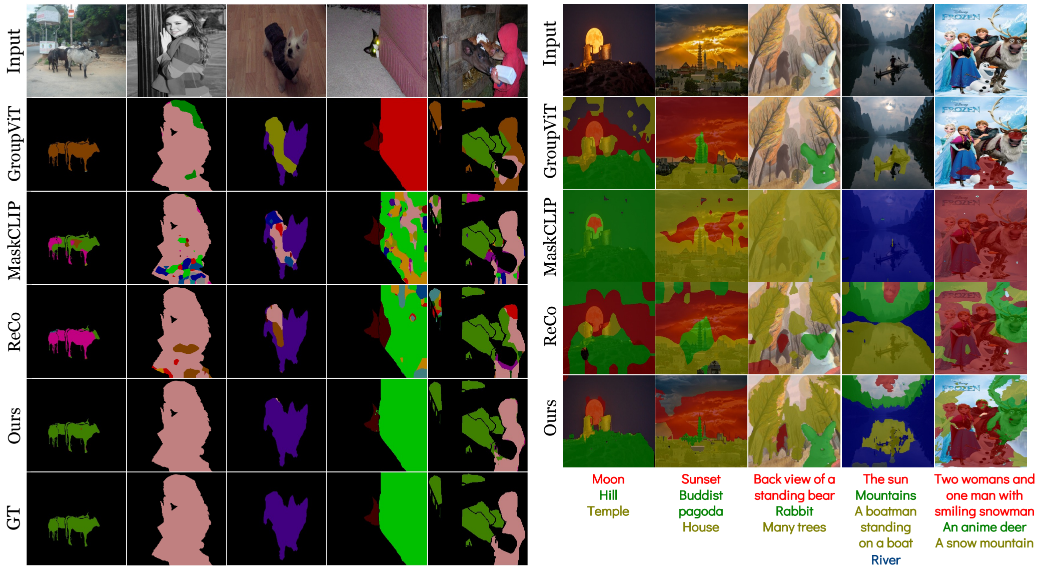
Even though the performance gap varies according to the characteristics of the evaluation dataset, GCL achieves state-of-the-art performance with a large margin in all datasets. Since region-level alignment learning of GCL lets the model learn the capability to distinguish the background region in a data-driven manner, GCL achieves remarkably outperforms the CL-based existing methods in datasets without background class also, without any heuristical consideration. These results support that our proposed GCL framework eliminates the train-test discrepancy in the alignment level existing in CL and successfully learns the region-level alignment.

4.3 Qualitative Results

We compare the proposed method qualitatively in Figure 4. At examples on VOC20 in Figure 4a, we observe that CL-based competing methods suffer from mis-alignment issue of regions and target classes, where clustering-based method (GroupViT) shows less-noisy segmentation masks compared to MaskCLIP and ReCo. In contrast, GCL provides more accurate segmentation masks through the direct region-text alignment learning. In addition, we present examples in the wild to show open-world segmentation capability in Figure 4b. For this purpose, we obtain segmentation masks for visual concepts not included in segmentation datasets (*e.g.*, moon, sunset) or free-form texts (*e.g.*, “two women and one man with a smiling snowman” and “a boatman standing on a boat”). GroupViT tends to focus on the main object of the image and regard the other objects as background, which is consistent with its good performance in object-oriented datasets. Interestingly, in contrast to the quantitative evaluation, we observe ReCo consistently outperforms MaskCLIP qualitatively. We conjecture that this is due to their different refinement approaches; the refinement approach of ReCo, retrieval and co-segmentation, is data-driven method while the refinement approach of MaskCLIP depends on heuristic post-processing which does not guarantee the general improvement. Finally, our method GCL consistently shows better results than other methods.

4.4 Ablation Study

We perform several ablation studies on PASCAL VOC and PASCAL Context to investigate the contributions of individual loss terms. For the ablation, we use a short learning schedule with a batch size of 512 and learning rate of 7.5×10^{-5}. for total $40,000$ iterations including $10,000$ warmup steps. Table 2a shows that the proposed GCL loss remarkably improves the segmentation performance ($26.9 \rightarrow 37.7$). Image-level or feature-level GCL loss solely improves the performance significantly, while using both losses together provides further performance gain. Also, we observe



(a) **Examples in VOC20 dataset.**

(b) **Examples in the wild.**

Figure 4: **Qualitative examples.** (a) The comparison shows the error types of each method in VOC dataset. The results only show a foreground region following the evaluation protocol. GroupViT clusters similar semantics well, but it occasionally causes incorrect segmentation of a large region. MaskCLIP tends to generate noisy mask. ReCo is less noisy, but still noisier than GroupViT or GCL. (b) shows results on the wild web images and free-form texts. Texts used as target classes are shown in the bottom of the images. This setting is more pertinent than evaluating on datasets of fixed classes for showing the open-world segmentation capability.

Table 2: **Ablation studies on GCL and prior losses.** We use short learning schedule for this ablation and refinement techniques are not applied to reveal the effect of each loss function clearly.

(a) **GCL loss** improves performance remarkably.

	GCL_v	GCL_f	CL	VOC	Context	Avg.
CL			✔	39.6	14.2	26.9
GCL$_v$	✔			50.8	19.1	34.9
GCL$_f$		✔		48.2	20.2	34.2
GCL	✔	✔		52.1	23.2	37.7
CL+GCL	✔	✔	✔	51.0	23.6	37.3

(b) **Prior losses.**

	$\mathcal{L}_{\text{area}}$	$\mathcal{L}_{\text{tv}}$	VOC	Context	Avg.
GCL	✔	✔	52.1	23.2	37.7
GCL$-\mathcal{L}_{\text{tv}}$	✔		47.5	21.5	34.5
GCL$-\mathcal{L}_{\text{area}}$		✔	28.7	22.1	25.4
GCL$-\mathcal{L}_{\text{tv}}-\mathcal{L}_{\text{area}}$			25.7	21.2	23.4

that incorporating CL with GCL is not effective. Table 2b shows the importance of prior losses. As described in the method section, area prior loss plays important role to prevent the model falling into a trivial solution.

5 Conclusion

We propose a novel framework for open-world semantic segmentation, which mainly tackles an issue in contrastive learning based methods—the discrepancy of the alignment level in training (*i.e.*, image-text) and inference (*i.e.*, region-text) time. In the framework, we incorporate the grounding process within contrastive learning, thus allowing explicit learning alignment between text and text-grounded regions (*i.e.*, segmentation result). We also present a unified evaluation protocol for a fair and direct comparison of existing methods. Finally, our GCL achieves state-of-the-art zero-shot segmentation performance on 8 datasets with large margins using the proposed protocol.

REFERENCES

Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4981–4990, 2018.

Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4253–4262, 2020.

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

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset, 2022.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1209–1218, 2018.

Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8991–9000, 2020.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568, 2021.

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. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017.

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

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

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

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *EMNLP*, 2021.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011.

Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5267–5276, 2019.

Peike Li, Yunchao Wei, and Yi Yang. Consistent structural relation learning for zero-shot segmentation. *Advances in Neural Information Processing Systems*, 2020.

Quande Liu, Youpeng Wen, Jianhua Han, Chunjing Xu, Hang Xu, and Xiaodan Liang. Open-world semantic segmentation via contrasting and clustering vision-language embedding. *European Conference on Computer Vision (ECCV)*, 2022.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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*, pp. 8748–8763. PMLR, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

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

Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. *Advances in Neural Information Processing Systems*, 2022.

Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7262–7272, 2021.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18134–18144, 2022.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. *CVPR*, 2022.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. *European Conference on Computer Vision (ECCV)*, 2022.

A Architecture Details

Our core design principle is to preserve and exploit the diverse knowledge of pre-trained CLIP. Therefore, we freeze the pre-trained CLIP and train the adapter network for the adaptation from the image-text alignment to the region-text alignment. We follow the simple modification of MaskCLIP (Zhou et al., 2022) to the image encoder. They modify the last attention of the CLIP image encoder to acquire the dense embedding representing local semantics. This dense image embedding is fed to the adapter. As shown in Fig 2, the adapter consists of four gated convolution blocks, where the output of convolution is gated by learned gating parameter and added to the skip connection. Concretely, the process of gated convolution can be written as:

$$\mathbf{x}' = \mathbf{x} + \text{tanh}(g) \cdot \text{Conv}(\mathbf{x}) \tag{13}$$

where $\mathbf{x}$ is input feature and g is a learned gating parameter. Furthermore, we employ two branch strategy. In addition to the main adapter branch, we use one more branch, named knowledge preservation branch. There are no learnable parameters in this branch and just mixed with the output of the adapter branch as following:

$$\mathbf{M}' = (1 - w_{kp}) \cdot \mathbf{M} + w_{kp} \cdot \mathbf{M}^{\text{KP}} \tag{14}$$

where $\mathbf{M}^{\text{KP}}$ is the generated masks of knowledge preservation branch, $\mathbf{M}'$ is the final output mask, and w_{kp} is a mixing hyperparameter. We use the w_{kp} of 0.3. To fully leverage the pre-trained knowledge, this branch is only used in the inference stage.

B Comparison with Zero-shot Segmentation Methods

Table 3: Zero-shot segmentation results for partial classes.

	VOC20	Context59	COCO-Stuff	Avg.
ViL-Seg	34.4	16.3	16.4	22.4
ReCo	57.9	32.0	18.4	36.1
GroupViT (RedCaps)	79.0	49.2	16.1	48.1
MaskCLIP	73.0	56.5	21.9	50.5
GCL (Ours)	**84.5**	**62.0**	**27.6**	**58.0**

We provide an extensive and unified comparison in Section 4, but the comparison cannot include a method that is not open-sourced, *e.g.*, ViL-Seg (Liu et al., 2022). They show the comparison in the conventional zero-shot segmentation protocol, where only partial classes are used for evaluation: 5 classes (potted plant, sheep, sofa, train, tv-monitor) for PASCAL VOC20, 4 classes (cow, motorbike, sofa, cat) for PASCAL Context59, and 15 classes (frisbee, skateboard, cardboard, carrot, scissors, suitcase, giraffe, cow, road, wall concrete, tree, grass, river, clouds, playingfield) for COCO-Stuff datasets. Therefore, we additionally provide the comparison results under the partial classes protocol. As shown in Table 3, GCL achieves state-of-the-art performance with a large margin in every dataset again.