CLIM: Contrastive Language-Image Mosaic for Region Representation

Size Wu[1], Wenwei Zhang[1], Lumin Xu[2]
Sheng Jin[3,4], Wentao Liu[4,5], Chen Change Loy[1]*
[1]S-Lab, Nanyang Technological University
[2]The Chinese University of Hong Kong
[3]The University of Hong Kong
[4]SenseTime Research and Tetras.AI
[5]Shanghai AI Laboratory
{size001, wenwei001, ccloy}@ntu.edu.sg, luminxu@link.cuhk.edu.hk, {jinsheng, liuwentao}@tetras.ai

Abstract

Detecting objects accurately from a large or open vocabulary necessitates the vision-language alignment on region representations. However, learning such a region-text alignment by obtaining high-quality box annotations with text labels or descriptions is expensive and infeasible. In contrast, collecting image-text pairs is simpler but lacks precise object location information to associate regions with texts. In this paper, we propose a novel approach called Contrastive Language-Image Mosaic (CLIM), which leverages large-scale image-text pairs effectively for aligning region and text representations. CLIM combines multiple images into a mosaicked image and treats each image as a *'pseudo region'*. The feature of each pseudo region is extracted and trained to be similar to the corresponding text embedding while dissimilar from others by a contrastive loss, enabling the model to learn the region-text alignment without costly box annotations. As a generally applicable approach, CLIM consistently improves different open-vocabulary object detection methods that use caption supervision. Furthermore, CLIM can effectively enhance the region representation of vision-language models, thus providing stronger backbones for open-vocabulary object detectors. Our experimental results demonstrate that CLIM improves different baseline open-vocabulary object detectors by a large margin on both OV-COCO and OV-LVIS benchmarks. The code is available at https://github.com/wusize/CLIM.

Introduction

Object detection is a fundamental task in computer vision that involves recognizing and localizing objects in the images. With the advent of deep learning, significant progress has been made in object detection on benchmark datasets that involve a confined set of categories, such as 80 classes in COCO (Lin et al. 2014) and 20 classes in PASCAL VOC (Everingham et al. 2010). However, to make object detection useful in real-world applications, it is essential to recognize objects in an open vocabulary that is inherently long-tailed and open-ended, with novel concepts that are not commonly seen in the benchmark datasets (Liu et al. 2019; Reed 2001).

To obtain the generalization ability required by open-vocabulary recognition, modern open-vocabulary object detection methods either apply the image-text pairs as weak supervision to the training of object detection (Zhou et al. 2022) or reap the vision-language alignment from large-scale image-text pre-training (Gu et al. 2021; Kuo et al. 2022; Wu et al. 2023b), which can be summarized as Figure 1(a). Due to lack of object location information, these methods fail to effectively transfer the vision-language alignment to region representations.

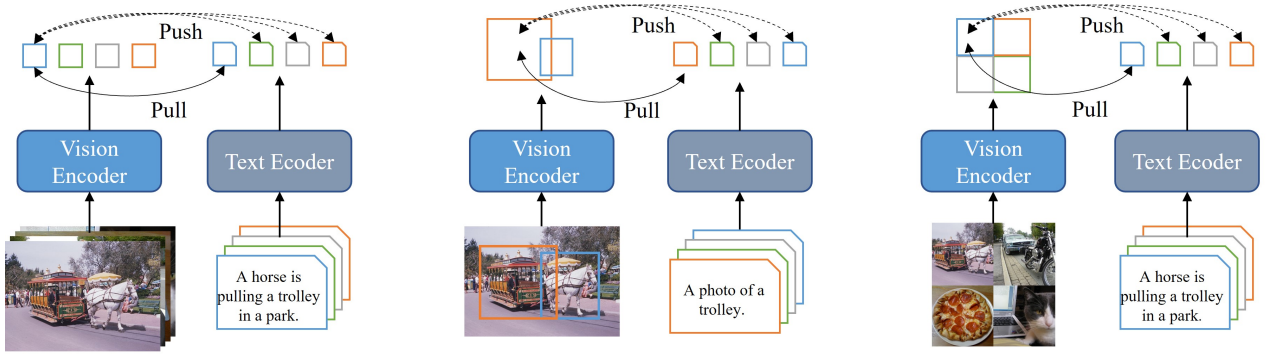
There are also attempts (Zhong et al. 2022; Lin et al. 2023) to learn vision-language alignment directly at the region level as shown in Figure 1(b). For example, RegionCLIP (Zhong et al. 2022) matches region proposals with object nouns to generate region-text pairs. The region proposals are detected by a pre-trained Region Proposal Network (RPN) and the object nouns are obtained by parsing the image captions. However, the generated region-text pairs are inevitably noisy since both the localization of region proposals and the region-text matching can be inaccurate.

In this paper, we propose a novel approach to learn region-language alignment without the inaccurate and tedious region-text matching process. The proposed method, named Contrastive Language-Image Mosaic (CLIM), combines multiple images into a mosaicked image and conducts contrastive learning as shown in Figure 1(c). This process forces the representation of each sub-image in the mosaicked image to be similar to its corresponding text representation and dissimilar to the others. By treating the sub-images as *'pseudo regions'*, CLIM facilitates the learning of region-text alignment while eliminating the need to annotate bounding boxes of objects.

CLIM prepares a canvas at each training iteration and divides it evenly into a flexible number of regions (*e.g.*, 2×2, 3×3 and 4×4). Each region is then filled by an image that is randomly sampled from the training dataset, termed as a *'pseudo region'*. The canvas then becomes a mosaicked image, which is fed into the vision encoder as a whole to obtain a feature map. The feature of each pseudo region is extracted from the feature map using the corresponding box location in the mosaicked image. Meanwhile, the texts of the original images are fed independently and parallelly to the text encoder. Finally, the features of pseudo regions are aligned with the text features in a contrastive manner.

We deliberately keep the design of CLIM simple so that it can be easily applied to different open-vocabulary object de-

*Corresponding author.



(a) Using Image-Text Pairs (b) Using Noisy Region-Text Pairs (c) Using Pseudo Region-Text Pairs

Figure 1: Different strategies of learning vision-language alignment. (a) Learning image-text alignment using image-text pairs. (b) Learning region-text alignment using noisy and un-scalable region-text pairs. (c) The proposed CLIM method mosaics images to generate pseudo region-text pairs for region-text alignment.

tection methods (Zhou et al. 2022; Wu et al. 2023b) as well as the pre-training of vision-language models (Radford et al. 2021; Zareian et al. 2021). To evaluate the effectiveness of CLIM, we test different design choices for both types of applications. When applied to the caption supervision branch of Detic (Zhou et al. 2022) and BARON (Wu et al. 2023b), CLIM improves Detic by 5.1 AP_{50}^{novel} and BARON by 2.1 AP_{50}^{novel} on OV-COCO. On OV-LVIS, it improves Detic by 2.3 mAP_{r}^{mask}. For the enhancement of CLIP models' region representation, the model trained by CLIM improves F-VLM (Kuo et al. 2022) by 2.2 mAP_{r}^{mask} on OV-LVIS. CLIM also boosts OV-RCNN (Zareian et al. 2021) on OV-COCO by 3.4 AP_{50}^{novel} when applied to OV-RCNN's vision-language pre-training stage.

Related Work

Learning Vision-Language Alignment. Aligning visual and lingual representations is a key step to achieve general scene understanding (Radford et al. 2021; Jia et al. 2021; Zareian et al. 2021; Kim, Son, and Kim 2021; Wu et al. 2023a). CLIP models (Radford et al. 2021) that are pre-trained on billion-scale image-text pairs have shown impressive zero-shot capabilities in downstream image recognition tasks. However, CLIP models lack awareness of local regions while building image-level alignment. To reason about image regions, RegionCLIP (Zhong et al. 2022) fine-tunes CLIP models by exploiting the region-text correspondence between pseudo-labeled region proposals and object concepts. On the other hand, GLIP (Li* et al. 2022) exploits visual-grounding datasets (Krishna et al. 2017; Hudson and Manning 2019) that associate bounding boxes and text descriptions. Nonetheless, datasets with bounding box annotations needed by these methods are limited and expensive, while large-scale image-text pairs lack object location information, preventing more assertive exploration of vision-language alignment at the region level. In this paper, we circumvent this issue by mosaicking images and treating each image as a 'pseudo region' for learning region-level representations using low-cost image-text annotations.

Weakly-Supervised Object Detection. The task of weakly-supervised object detection (WSOD) is to train detectors using image-level supervision. Some studies (Li et al. 2019; Shen et al. 2019; Wan et al. 2019) rely on low-level region proposal techniques (Uijlings et al. 2013; Arbeláez et al. 2014) to localize objects. A more general form of WSOD, known as semi-supervised WSOD (Zhang et al. 2021; Ramanathan, Wang, and Mahajan 2020; Redmon and Farhadi 2017), allows the use of bounding box supervision together with image labels. In particular, MosaicOS (Zhang et al. 2021) groups object-centric images into pseudo scene-centric images using mosaic augmentation to address the challenge of long-tail object detection. CLIM is similar to MosaicOS in mosaicking images. However, MosaicOS serves as an augmentation to re-balance the distributions of rare and frequent categories while CLIM aims at the representation learning for region-language alignment. Moreover, MosaicOS is specially designed for object detection while CLIM is a more generally applicable approach that applies to not only open-vocabulary detection methods but also vision-language pre-training.

Open-Vocabulary Object Detection. Open-vocabulary object detection (OVD) (Zareian et al. 2021) is concerned with detecting objects of novel categories that are unseen during training. Some works utilize large pre-trained vision-language models (VLMs) (Radford et al. 2021) to acquire open-vocabulary recognition ability by knowledge distillation (Gu et al. 2021; Zang et al. 2022; Wu et al. 2023b,c) or directly building open-vocabulary detectors upon the pre-trained VLMs (Zareian et al. 2021; Kuo et al. 2022; Xu et al. 2023). Others (Gao et al. 2021; Zhou et al. 2022; Wu et al. 2023b) employ image-level supervision (*e.g.*, image captions) to learn a large number of novel concepts. In this paper, we first instantiate CLIM on the works that use image captions, *i.e.*, Detic (Zhou et al. 2022) and BARON (Wu et al. 2023b). Then, we apply CLIM to the vision-language pre-training so that the VLMs serve as stronger backbones to build open-vocabulary object detectors for F-VLM (Kuo et al. 2022) and OV-RCNN (Zareian et al. 2021).

Method

In this section, we introduce CLIM that facilitates the learning of region-level visual-language alignment from image-text pairs, without relying on either costly bounding box annotations (Li et al. 2022) or inaccurate bounding box predictions (Zhou et al. 2022; Lin et al. 2023). Our method involves mosaicking images and treating each image as a *'pseudo region'* in the context of the mosaicked image. The visual features of these pseudo regions are then extracted and aligned with their corresponding text features through contrastive learning. CLIM is versatile and can be applied to both open-vocabulary object detection methods and vision-language pre-training.

Mosaicking Images as Pseudo Regions

Ideally, learning the alignment between region and text representations for generalizable object recognition would require one to annotate massive region-level bounding boxes that are labeled with text descriptions, which can be costly and time-consuming. CLIM offers a solution that uses image-text pairs by mosaicking multiple images and treating each image as a *'pseudo region'*. Through this approach, we can obtain an accurate mapping between the pseudo regions and their corresponding text descriptions in the mosaicked image 'for free' (at a much lower cost than box labeling).

At each training iteration, we begin by preparing a large canvas that is equally divided into square regions, such as 2×2, 3×3 or 4×4 grids. Next, we sample 4, 9 or 16 image-text pairs from the training dataset, and each image is randomly cropped and then resized to fill a unique square region. In this way, the box location of a pseudo region in the mosaicked image and its corresponding region description are obtained without manual annotations, making it possible to train region-level alignment via contrastive learning.

Aligning Region and Text Representations

Given a mosaicked image that contains several pseudo regions and their corresponding text descriptions, CLIM aligns the region features with their corresponding text descriptions. As shown in Figure 2, the mosaicked image is first sent to a vision encoder, which outputs the feature map of the mosaicked image. Then the feature of each pseudo region f_v can be extracted according to the corresponding location. For the language representation, the text description of each pseudo region is separately fed to the text encoder to obtain the text embedding f_t.

Given region features f_v and their corresponding text embeddings f_t, we conduct contrastive learning to force region features to be similar to their text features and dissimilar to those of others. Specifically, the model is trained to maximize the cosine similarity of matched pairs $\langle f_v^+, f_t^+ \rangle$ and minimize the cosine similarity of unmatched pairs $\langle f_v^+, f_t^- \rangle$.

Applications of CLIM

We apply CLIM to the fine-tuning stage of Detic (Zhou et al. 2022) and BARON (Wu et al. 2023b) for open-vocabulary object detection (OVD). Besides, we use CLIM to enhance CLIP (Radford et al. 2021) model's region representation, and also equip OV-RCNN (Zareian et al. 2021) with CLIM in its vision-language pre-training stage.

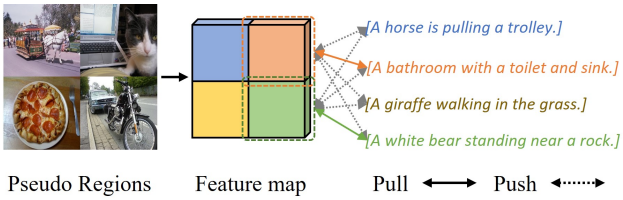


Figure 2: An overview of CLIM. Features of pseudo regions are learned to be similar to the corresponding text embeddings (the colored arrows) and dissimilar to uncorresponding ones (the grey arrows). This process can be applied to both the open-vocabulary object detection (*e.g.*, Detic) and the the pre-training of vision-language models (*e.g.*, CLIP).

OVD with Detic. Detic (Zhou et al. 2022) adopts a two-stage training for open-vocabulary detection. It first trains a detector on base categories, and then fine-tunes the detector using image-level supervision. During the fine-tuning stage, it aligns region feature of the max-size proposal to text embedding of the image label, or aligns the feature of the image box to the text embedding of the caption. We apply CLIM to Detic by mosaicking images first and aligning the visual features of pseudo regions to the corresponding text embeddings. As for the box location for region feature extraction, we follow the original Detic to use the box enclosing the pseudo region for caption loss and use the max-size box inside the pseudo region for image tag (label) loss.

OVD with BARON. BARON (Wu et al. 2023b) represents an image using a bag of regions. It samples several region proposals in an image and projects region features into word embedding space (pseudo words). Then the pseudo words are concatenated and sent to the text encoder for bag-of-regions embedding. The bag-of-regions embeddings are aligned to the caption embeddings of corresponding images. When applying CLIM, we sample region proposals inside a pseudo region (sub-image), and use the bag-of-regions embedding to represent the pseudo region, which is aligned to the corresponding caption embedding using the contrastive loss (Yang et al. 2022) employed in BARON.

Enhancing CLIP Model's Region Representation. Contrastive Language-Image Pretraining (CLIP) (Radford et al. 2021) learns vision-language alignment on large-scale paired images and texts (Schuhmann et al. 2022; Sharma et al. 2018). For an image-text pair, it obtains a global representation of the image using a [CLS] token in the attention module. CLIM concerns region representations instead of a global representation for the downstream dense prediction applications (*i.e.*, object detection). Therefore, we follow the practice of MaskCLIP (Zhou, Loy, and Dai 2022) to extract a feature map from the last attention layer of CLIP's vision model, after which we apply RoIAlign (He et al. 2017) on the feature map to extract the region representations. Then we use the same contrastive loss (*i.e.*, InfoNCE

loss) in CLIP (Radford et al. 2021) to align region representations and the corresponding text representations. The presence of region-level supervision would significantly enhance the CLIP model's region-text alignment.

Vision-language Pre-training Stage of OV-RCNN. OV-RCNN (Zareian et al. 2021) pre-trains a vision backbone and a vision-to-language projection layer that maps region features into word embedding space. It splits an image into N grid regions and calculates the similarity between N grid embeddings and C word embeddings of the caption text. The similarity score of the image-text pair is obtained by averaging the $N \times C$ grid-word cosine similarities, and supervised by a grounding loss (CE loss) that maximizes similarities of matched image-text pairs and minimizes that of unmatched pairs. When applying CLIM, we divide each pseudo region into grids and average the grid-word similarities to obtain the similarity score between the pseudo region and a text. The same grounding loss is used to impose the alignment between pseudo regions and texts.

Discussion

CLIM v.s. Mosaic Augmentation. CLIM is reminiscent of the Mosaic augmentation (Bochkovskiy, Wang, and Liao 2020) typically used in conventional object detection tasks, where Mosaic helps manipulate data distribution, increase data diversity and reduce the need for a large batch size. However, CLIM is distinguished from such data augmentation techniques as it is intrinsically a representation learning paradigm that targets at region-language alignment. Data augmentations concern manipulating images in the preprocessing stage, while CLIM focuses on the contrastive learning between pseudo regions and corresponding texts during model training to strengthen the region recognition ability. Besides, the data augmentations are specially designed for object detection and require box annotations. On the contrary, CLIM treats images as regions and generally applies to not only open-vocabulary detection methods but also vision-language pre-training.

Comparison with Manual Labelling. Compared with CLIM, manually labeling region-text pairs is quite labor-intensive and inefficient. For example, for 4,000 referring expressions of regions, the labeling needs 3 weeks of crowdsourcing as reported in RefCOCO (Kazemzadeh et al. 2014), which wrapped the labelling process as an online computer game and collected the players' annotations from the web. In contrast, CLIM circumvents such region-level annotation by generating pseudo region-text pairs.

Experiments

In the experiment section, we first introduce the main datasets and evaluation metrics. Then we separately introduce the applications to OVD methods (*i.e.*, Detic and BARON) and vision-language pre-training (*i.e.*, CLIP and OV-RCNN) with their implementation details and ablation study on the design choices.

Datasets and Evaluation Metrics

We focus on open-vocabulary object detection and report results on the OV-COCO and OV-LVIS benchmarks.

OV-COCO. We follow OV-RCNN (Zareian et al. 2021) to divide COCO dataset (Lin et al. 2014) into 48 base classes and 17 novel classes. The training set contains 107,761 images of base category annotations, and the test set contains 4,836 images with both base and novel category annotations. We report the box AP at IoU threshold 0.5, which is denoted as AP_{50}. AP_{50} of novel categories ($\text{AP}_{50}^{\text{novel}}$) is the major metric to evaluate the OVD performance on OV-COCO benchmark.

OV-LVIS. We follow ViLD (Gu et al. 2021) to use the 866 common and frequent classes in LVIS dataset (Gupta, Dollár, and Girshick 2019) as base categories, and 337 rare classes as novel categories. Only base category annotations are used during training. For OV-LVIS, we report mean Average Precision (mAP) of masks averaged on IoUs from 0.5 to 0.95. The mAP of rare categories ($\text{mAP}_r^{\text{mask}}$) is the main evaluation metric for OV-LVIS benchmark.

Application to Detic & BARON

We apply CLIM to the OVD methods that use caption supervision, *i.e.*, Detic (Zhou et al. 2022) and BARON (Wu et al. 2023b). The ablation studies are conducted on Detic using caption loss only on the OV-COCO benchmark as shown in Table 2.

Implementation Details. For Detic (Zhou et al. 2022), we use the Faster RCNN with ResNetC4 (Ren et al. 2015) backbone as the detector on OV-COCO benchmark, and use the detector based on CenterNet2 (Zhou, Koltun, and Krähenbühl 2021) on OV-LVIS benchmark. For the caption supervision, COCO Caption (Chen et al. 2015) is used on OV-COCO benchmark and CC3M (Sharma et al. 2018) is used on OV-LVIS benchmark. Detic (Zhou et al. 2022) alternatively applies image-level supervision and box-level supervision during the fine-tuning stage in its official implementation. We re-implement it by applying the two sources of supervision in parallel and thus requiring only half of the total iterations (45,000). We also re-implement BARON (Wu et al. 2023b) in the same way by adopting the two-stage training and parallel supervision. Our re-implementation consistently outperforms the results reported in the original papers of Detic (Zhou et al. 2022) and BARON (Wu et al. 2023b).

OV-COCO. In Table 1a, we report the results of applying CLIM to Detic and BARON on OV-COCO benchmark. When applied to the image-level supervision of Detic and BARON at the fine-tuning stage, CLIM increases the performance on novel categories by 5.1 AP_{50} and 2.1 AP_{50}, respectively.

OV-LVIS. On OV-LVIS benchmark, we apply CLIM to Detic (Zhou et al. 2022) and its follow-up VLDet (Lin et al. 2023). As shown in Table 1b, CLIM improves Detic by 2.3 $\text{mAP}_r^{\text{mask}}$. CLIM also improves VLDet that is built upon Detic and matches region proposals with object concepts to

Method	Backbone	AP_{50}^{novel}
ViLD (Gu et al. 2021)	RN50	27.6
OV-DETR (Zang et al. 2022)	RN50	29.4
RegionCLIP (Zhong et al. 2022)	RN50	26.8
PB-OVD (Gao et al. 2021)	RN50	30.8
VLDet (Lin et al. 2023)	RN50	32.0
F-VLM (Kuo et al. 2022)	RN50	28.0
OADP (Wang et al. 2023)	RN50	35.6
OV-RCNN (Zareian et al. 2021)	RN50	22.8
Detic* (Zhou et al. 2022)	RN50	30.3
BARON* (Wu et al. 2023b)	RN50	34.8
OV-RCNN (Zareian et al. 2021) + CLIM	RN50	26.2(+3.4)
Detic* (Zhou et al. 2022) + CLIM	RN50	35.4(+5.1)
BARON* (Wu et al. 2023b) + CLIM	RN50	36.9(+2.1)

(a) OV-COCO benchmark

Method	Backbone	mAP_r^{mask}
ViLD (Gu et al. 2021)	RN50	16.6
OV-DETR (Zang et al. 2022)	RN50	17.4
DetPro (Du et al. 2022)	RN50	19.8
OC-OVD (Rasheed et al. 2022)	RN50	21.1
BARON-KD (Wu et al. 2023b)	RN50	22.6
RegionCLIP (Zhong et al. 2022)	RN50	17.1
OADP (Wang et al. 2023)	RN50	21.7
Detic (Zhou et al. 2022)	RN50	19.5
VLDet (Lin et al. 2023)	RN50	21.7
F-VLM (Kuo et al. 2022)*	RN50x64	30.1
Detic (Zhou et al. 2022) + CLIM	RN50	21.8(+2.3)
VLDet (Zhou et al. 2022) + CLIM	RN50	22.2(+0.5)
F-VLM (Kuo et al. 2022)* + CLIM	RN50x64	32.3(+2.2)

(b) OV-LVIS benchmark

Table 1: Results on open-vocabulary object detection. * means the marked methods are reproduced by us.

#	Batch Size	CLIM	AP_{50}^{novel}
1	4	✗	24.4
2	20	✗	27.4
3	32	✗	26.3
4	≈ 20	✓	32.3

(a) Sanity check.

#	Grid Size	AP_{50}^{novel}
1	2×2	30.9
2	3×3	31.2
3	4×4	30.8
4	random	32.3

(b) Number of pseudo regions.

#	Sampling	AP_{50}^{novel}
1	text	31.3
2	image	31.6
3	random	32.3

(c) Sampling of pseudo regions.

#	CLIM	Tag loss	AP_{50}^{novel}
1	✗	✗	24.4
2	✗	✓	30.3
3	✓	✗	32.3
4	✓	✓	35.4

(d) Using image tag loss.

Table 2: Ablation study of components in CLIM on Detic.

produce region-text pairs. This indicates that CLIM is orthogonal to methods that adopt pseudo-labelling strategies to obtain object-language correspondence.

Sanity Check. We verify that the improvement of CLIM is not mainly attributed to the increased number of images. In our implementation of Detic, we set the batch size of box and caption supervision as 2 and 4 on each GPU, respectively. When applying CLIM to Detic, we further split the 4 images for caption supervision into 2 with mosaic and 2 without mosaic. And we randomly choose 2×2, 3×3 and 4×4 mosaic. Therefore, there are on average $2 + 2 \times (4 + 9 + 16)/3 \approx 20$ images for caption supervision in each iteration. As the sanity check, we replace CLIM with simply stacking more images for caption supervision in a batch. As shown in Table 2a, increasing the batch size of caption supervision from 4 to 20 only leads to 3.0 performance gain on novel categories, while CLIM achieves 7.9 performance gain (32.3 AP_{50}^{novel} v.s. 24.4 AP_{50}^{novel}) under this fair comparison. Besides, we also observe that further increasing the batch size does not bring any improvement in Table 2a (#4).

Number of Pseudo Regions. In Table 2b, we study the number of pseudo regions in a mosaicked image. The overall resolution of the mosaicked image is fixed as 800×800 in this ablation study. We first separately choose 2×2, 3×3 and 4×4 and observe that the 3×3 mosaic achieves the best performance on novel categories (31.2 AP_{50}^{novel}). However, the performance gap between these single-pattern choices is marginal. When we randomly choose from the three settings in each iteration, the AP_{50} on novel categories grows by 1.1. This indicates that the mixed use of 2×2, 3×3 and 4×4 mosaic allows the model to generalize to different region patterns, thus improving the performance on novel categories. However, simply increasing the number of pseudo regions does not provide consistent performance gain (#2 and #3).

Sampling of Pseudo Regions. By default, we randomly combine pseudo regions. And we also consider grouping samples that are similar in CLIP text representations or image representations by calculating cosine distance. However, both approaches decrease the performance as shown in Table 2c. Combining pseudo regions that have similar text descriptions or visual contents tends to generate mosaiked images of a specific pattern, limiting the models' generalization ability to different region patterns in testing.

Image Tag Loss in Detic. Detic can be trained with caption loss only, or with both max-size image tag (label) loss and the caption loss. We report the results before and after adding image tag loss in Table 2d. We show that the image tag loss improves the baseline to 30.3 AP_{50}^{novel} (#1 and #2), and also improves the performance of our CLIM to 35.4 AP_{50}^{novel} (#3 and #4). This also indicates that CLIM can bring consistent performance gain on different variants of Detic.

Enhancing CLIP's Region Representation

We study the application of CLIM to CLIP models (Radford et al. 2021) and analyze how CLIM would help improve CLIP's region representation. For simplicity, we use the Top-1 and Top-5 accuracy of classifying the COCO dataset's bounding boxes to evaluate the enhancement of region representation. In addition to the zero-shot region classification, we also build open-vocabulary detectors upon the models trained by CLIM following F-VLM (Kuo et al. 2022)

#	Data	Resolution	Top1	Top5
1	COCO Caption	-	29.2	51.6
2	COCO Caption	320×320	57.8	80.0
3	COCO Caption	640×640	61.3	83.8
4	COCO Caption	1024×1024	62.2	84.3

Table 3: Resolution of the mosaicked images.

to further validate the enhancement of region representation for realistic downstream application.

Implementation Details. When applying CLIM to CLIP models, we mainly study the ViT-B-16 variant of CLIP and use the model weights released by OpenAI to initialize our training. For the experiment on OV-COCO, we train the CLIP model on COCO Caption (Chen et al. 2015) for 100 epochs. For the experiment on OV-LVIS, we train the CLIP model on CC3M (Sharma et al. 2018) for 3 epochs. Following the pre-training of the original CLIP models (Radford et al. 2021), we use AdamW optimizer and set the batch size to 128 and the learning rate to 1e-5. After being trained by CLIM, the CLIP models are then used to build open-vocabulary detectors.

Image Resolution for CLIM. We experiment with different resolutions (320×320, 640×640, 1024×1024) of the mosaicked images to train the model. As shown in Table 3, CLIM significantly improves the region representation of CLIP, and increasing input resolution consistently produces performance gains. However, we do not further enlarge the input size and apply 1024×1024 as image resolution for training due to the quartically increasing computation cost.

Building Open-Vocabulary Object Detector. In addition to zero-shot inference on classifying ground truth bounding boxes, we also built open-vocabulary detectors following the architecture of F-VLM (Kuo et al. 2022), to verify the enhancement of region representation. As shown in Table 1b, with the RN50x64 model trained by CLIM, we improve F-VLM by 2.2 $\text{mAP}_r^{\text{mask}}$ on the OV-LVIS benchmark. Besides, we also build F-VLM with the ViT-B-16 model on the OV-COCO and OV-LVIS benchmarks as shown in Table 4. To obtain multi-scale feature maps for the Feature Pyramid Network (FPN) in the ViT-based detector, we extract feature maps from the 3th, 5th, 7th and 11th attention layers of the ViT model, and interpolate them to $[\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}]$ of the input image size. We observe weak performances of the ViT-based detector on both benchmarks for both base and novel categories when using the original ViT-B-16 model released by OpenAI, which are significantly improved by the model trained by CLIM (#2). Due to the lack of translation invariance and equivariance in the transformer architecture, the presence of region-level supervision by CLIM is particularly necessary during the pre-training of ViT-based vision-language models for downstream dense prediction tasks like object detection.

Comparison with RegionCLIP. We compare our approach with RegionCLIP (Zhong et al. 2022) in the ability of enhancing region representation in Table 5. RegionCLIP matches region proposals with object nouns to generate the region-text pairs for the learning of region-language alignment. We implement RegionCLIP using ViT-B-16 and COCO Caption dataset (Chen et al. 2015). The region proposals are detected by an region proposal network (RPN) trained on COCO's box annotations of base categories, which are cropped and sent to CLIP's image encoder (ViT-B-16) to obtain region representations. The cosine similarities between region representations and text representations of object nouns are used as the metric for matching. We also set the input image resolution as 1024×1024 to train RegionCLIP. It is noticeable that RegionCLIP is built on a strong assumption of the existence of a well-trained RPN and a pre-defined vocabulary of object nouns (Zhong et al. 2022). However, we still achieve comparable results on both zero-shot region classification and open-vocabulary object detection.

Model	OV-COCO		OV-LVIS		
	$\text{AP}_{50}^{\text{novel}}$	$\text{AP}_{50}^{\text{base}}$	$\text{mAP}_r^{\text{mask}}$	$\text{mAP}_c^{\text{mask}}$	$\text{mAP}_f^{\text{mask}}$
CLIP	21.6	36.4	14.8	20.5	26.1
CLIM	**25.7**	**42.5**	**20.8**	**25.6**	**29.7**

Table 4: Open-vocabulary detection results of applying CLIM to CLIP's ViT-B-16 model.

#	Method	Region Classification		OV-COCO	
		Top1	Top5	$\text{AP}_{50}^{\text{novel}}$	$\text{AP}_{50}^{\text{base}}$
1	CLIP	29.2	51.6	21.6	36.4
2	RegionCLIP	62.8	84.7	26.1	42.4
3	CLIM	62.2	84.3	25.7	42.5

Table 5: Comparison with RegionCLIP on zero-shot region classification and open-vocabulary object detection. RegionCLIP applies the well-trained RPN and pre-defined vocabulary of object nouns for region-text alignment.

Application to OV-RCNN

For OV-RCNN (Zareian et al. 2021), we follow the official implementation, which pre-trains the model for 40,000 iterations with a batch size of 64 on COCO Caption (Chen et al. 2015) and finetunes it for 150,000 iterations with a batch size of 8 on the box annotations of base categories in COCO dataset (Lin et al. 2014). Our CLIM is applied in the pre-training stage. As shown in Table 1a, CLIM boosts the final performance of OV-RCNN on novel categories by 3.4 AP_{50}.

Visualization & Analysis

We provide visualization and analysis of the enhancement of region representation in this section. First, we analyze the open-vocabulary object detector (Detic) trained with CLIM. Then we visualize how CLIM improves the vision-language alignment of CLIP's region representation.

Region Response to Text Description

As images are taken as pseudo regions during the training of CLIM, the models should have earned the generalization ability to localize regions given corresponding text de-

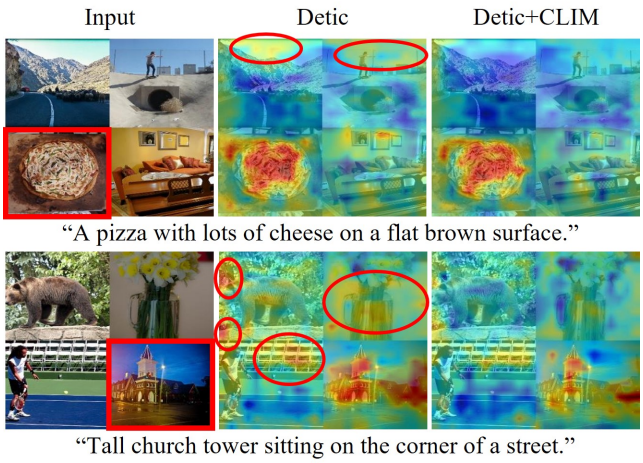


Figure 3: Feature map response on mosaicked images. The sub-images with a red border correspond to the text descriptions below. False positive regions with high response values are highlighted with red circles.

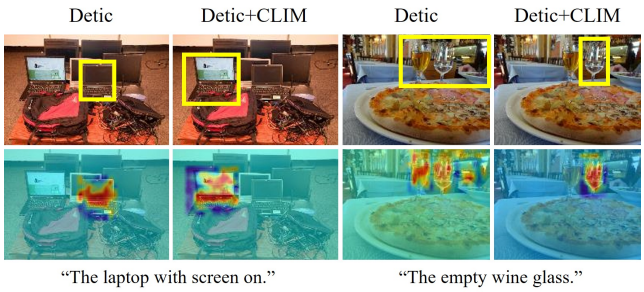


Figure 4: Feature map response on natural images. The yellow boxes are the detected bounding boxes of the queried text descriptions.

scriptions. Specifically, we analyze the models' ability to response to queried texts at the regions of interest by calculating cosine similarities between feature map and the text embeddings. We compare the features map response of the Detic baseline and the Detic trained with CLIM. And we consider two types of images, *i.e.*, mosaicked images that only appear during training and natural images that are ubiquitous in testing.

Mosaicked Images. As shown in Figure 3, the feature map response of our model is more concentrated on the pseudo region corresponding to the queried text. In comparison, the distribution of the Detic baseline's feature map response is more diffused. And there are many high responses outside of the queried pseudo regions as highlighted by red circles in Figure 3.

Natural Images. For the natural images, we not only visualize the feature map response but also show if the detector can localize the queried text by bounding boxes. This is similar to the task of referring expression comprehension (Yu et al. 2016), where the model is required to accurately locate the desired object given a text description that contains the attributes or the context of the object. As shown in Figure 4, the Detic baseline model detects the undesired objects or cannot accurately locate the objects. Although the Detic baseline model has learned the vision-language alignment at the image level, it cannot effectively transfer the alignment knowledge to regions. In comparison, our CLIM, which forms pseudo regions by mosaicking images, forces the detector to learn to associate regions with texts.

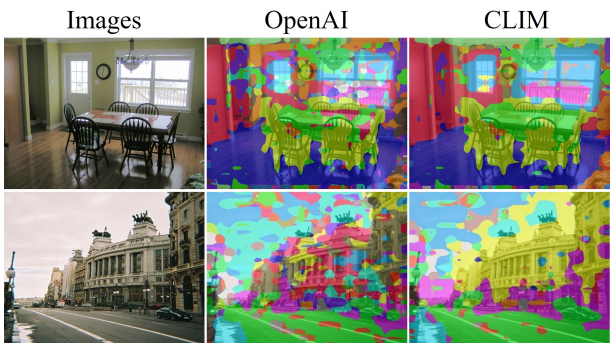


Figure 5: Visualization of CLIP's feature map by per-pixel classification. The images are from PASCAL Context (Everingham et al. 2010) dataset with 59 pixel categories.

Enhancement of CLIP's Region Representation

To verify CLIM's effectiveness in improving CLIP's region representation, we visualize the vision-language alignment of CLIP's feature map using per-pixel classification. Specifically, we choose the ViT-B-16 and compare the original model released by OpenAI with the model trained with CLIM. The images are from PASCAL Context dataset (Everingham et al. 2010) with 59 pixel categories. As shown in Figure 5, the feature map visualization of CLIM model is less diffused and different objects are more accurately divided, indicating the significant improvement of vision-language alignment on CLIP model's feature map. CLIM benefits per-pixel recognition even though the single pixel embeddings are not directly supervised during training.

Conclusion

In this paper, we present a novel method, Contrastive Language-Image Mosaic (CLIM), to exploit large-scale image-text pairs for region-language alignment without needing expensive bounding box annotations or relying on inaccurate box predictions. CLIM achieves this by mosaicking multiple images and treating each image as a *'pseudo region'* within the context of the mosaicked image, and then learning region-level representations via contrastive learning. By eliminating the need for costly annotations or noisy box predictions, CLIM presents an efficient and general solution with only image-text pairs for training open-vocabulary object detectors as well as improving vision-language model's region representation.

Acknowledgments

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-PhD-2023-08-048T), the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). Besides, we thank Dr. Xiangtai Li for his help in building detectors on CLIP models, *e.g.*, reproducing F-VLM.

References

Arbeláez, P.; Pont-Tuset, J.; Barron, J. T.; Marques, F.; and Malik, J. 2014. Multiscale combinatorial grouping. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 328–335.

Bochkovskiy, A.; Wang, C.-Y.; and Liao, H.-Y. M. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv: Computer Vision and Pattern Recognition*.

Du, Y.; Wei, F.; Zhang, Z.; Shi, M.; Gao, Y.; and Li, G. 2022. Learning To Prompt for Open-Vocabulary Object Detection With Vision-Language Model. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The PASCAL Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.*

Gao, M.; Xing, C.; Niebles, J. C.; Li, J.; Xu, R.; Liu, W.; and Xiong, C. 2021. Open Vocabulary Object Detection with Pseudo Bounding-Box Labels. *arXiv preprint arXiv:2111.09452*.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In *Int. Conf. Learn. Represent.*

Gupta, A.; Dollár, P.; and Girshick, R. B. 2019. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *Int. Conf. Comput. Vis.*

Hudson, D. A.; and Manning, C. D. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Int. Conf. Mach. Learn.*

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*.

Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *Int. Conf. Mach. Learn.*

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis.*

Kuo, W.; Cui, Y.; Gu, X.; Piergiovanni, A. J.; and Angelova, A. 2022. F-VLM: Open-Vocabulary Object Detection upon Frozen Vision and Language Models. *CoRR*, abs/2209.15639.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.; Chang, K.; and Gao, J. 2022. Grounded Language-Image Pre-training. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Li*, L. H.; Zhang*, P.; Zhang*, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; Chang, K.-W.; and Gao, J. 2022. Grounded Language-Image Pre-training. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Li, X.; Kan, M.; Shan, S.; and Chen, X. 2019. Weakly supervised object detection with segmentation collaboration. In *Int. Conf. Comput. Vis.*, 9735–9744.

Lin, C.; Sun, P.; Jiang, Y.; Luo, P.; Qu, L.; Haffari, G.; Yuan, Z.; and Cai, J. 2023. Learning Object-Language Alignments for Open-Vocabulary Object Detection. In *Int. Conf. Learn. Represent.*

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common objects in context. In *Eur. Conf. Comput. Vis.*

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019. Large-Scale Long-Tailed Recognition in an Open World. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*

Ramanathan, V.; Wang, R.; and Mahajan, D. 2020. Dlwl: Improving detection for lowshot classes with weakly labelled data. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Rasheed, H.; Maaz, M.; Khattak, M. U.; Khan, S.; and Khan, F. S. 2022. Bridging the Gap between Object and Image-level Representations for Open-Vocabulary Detection. In *NeurIPS*.

Redmon, J.; and Farhadi, A. 2017. YOLO9000: better, faster, stronger. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 7263–7271.

Reed, W. J. 2001. The Pareto, Zipf and other power laws. *Economics Letters*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Adv. Neural Inform. Process. Syst.*

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. *CoRR*, abs/2210.08402.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Shen, Y.; Ji, R.; Wang, Y.; Wu, Y.; and Cao, L. 2019. Cyclic guidance for weakly supervised joint detection and segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 697–707.

Uijlings, J. R.; Van De Sande, K. E.; Gevers, T.; and Smeulders, A. W. 2013. Selective search for object recognition. *Int. J. Comput. Vis.*, 104: 154–171.

Wan, F.; Liu, C.; Ke, W.; Ji, X.; Jiao, J.; and Ye, Q. 2019. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2199–2208.

Wang, L.; Liu, Y.; Du, P.; Ding, Z.; Liao, Y.; Qi, Q.; Chen, B.; and Liu, S. 2023. Object-Aware Distillation Pyramid for Open-Vocabulary Object Detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Wu, J.; Li, X.; Xu, S.; Yuan, H.; Ding, H.; Yang, Y.; Li, X.; Zhang, J.; Tong, Y.; Jiang, X.; Ghanem, B.; and Tao, D. 2023a. Towards Open Vocabulary Learning: A Survey. *arXiv pre-print*.

Wu, S.; Zhang, W.; Jin, S.; Liu, W.; and Loy, C. C. 2023b. Aligning Bag of Regions for Open-Vocabulary Object Detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Wu, S.; Zhang, W.; Xu, L.; Jin, S.; Li, X.; Liu, W.; and Loy, C. C. 2023c. CLIPSelf: Vision Transformer Distills Itself for Open-Vocabulary Dense Prediction. *arXiv preprint arXiv:2310.01403*.

Xu, S.; Li, X.; Wu, S.; Zhang, W.; Li, Y.; Cheng, G.; Tong, Y.; Chen, K.; and Loy, C. C. 2023. DST-Det: Simple Dynamic Self-Training for Open-Vocabulary Object Detection. *arXiv preprint arXiv:2310.01393*.

Yang, J.; Li, C.; Zhang, P.; Xiao, B.; Liu, C.; Yuan, L.; and Gao, J. 2022. Unified Contrastive Learning in Image-Text-Label Space. In *IEEE Conf. Comput. Vis. Pattern Recog.* IEEE.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *Eur. Conf. Comput. Vis.*

Zang, Y.; Li, W.; Zhou, K.; Huang, C.; and Loy, C. C. 2022. Open-Vocabulary DETR with Conditional Matching. *arXiv preprint arXiv:2203.11876*.

Zareian, A.; Rosa, K. D.; Hu, D. H.; and Chang, S.-F. 2021. Open-vocabulary object detection using captions. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Zhang, C.; Pan, T.-Y.; Li, Y.; Hu, H.; Xuan, D.; Changpinyo, S.; Gong, B.; and Chao, W.-L. 2021. MosaicOS: A Simple and Effective Use of Object-Centric Images for Long-Tailed Object Detection. In *Int. Conf. Comput. Vis.*

Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L. H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; et al. 2022. RegionCLIP: Region-based language-image pretraining. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Zhou, C.; Loy, C. C.; and Dai, B. 2022. Extract Free Dense Labels from CLIP. In *Eur. Conf. Comput. Vis.*

Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; and Misra, I. 2022. Detecting twenty-thousand classes using image-level supervision. In *Eur. Conf. Comput. Vis.*

Zhou, X.; Koltun, V.; and Krähenbühl, P. 2021. Probabilistic two-stage detection. In *arXiv preprint arXiv:2103.07461*.