Open-Vocabulary Panoptic Segmentation with MaskCLIP

Zheng Ding
UC San Diego
zhding@ucsd.edu

Jieke Wang
UC San Diego
jiw010@ucsd.edu

Zhuowen Tu
UC San Diego
ztu@ucsd.edu

Abstract

In this paper, we tackle a new computer vision task, open-vocabulary panoptic segmentation, that aims to perform panoptic segmentation (background semantic labeling + foreground instance segmentation) for arbitrary categories of text-based descriptions. We first build a baseline method without finetuning nor distillation to utilize the knowledge in the existing CLIP model. We then develop a new method, MaskCLIP, that is a Transformer-based approach using mask queries with the ViT-based CLIP backbone to perform semantic segmentation and object instance segmentation. Here we design a Relative Mask Attention (RMA) module to account for segmentations as additional tokens to the ViT CLIP model. MaskCLIP learns to efficiently and effectively utilize pre-trained dense/local CLIP features by avoiding the time-consuming operation to crop image patches and compute feature from an external CLIP image model. We obtain encouraging results for open-vocabulary panoptic segmentation and state-of-the-art results for open-vocabulary semantic segmentation on ADE20K and PASCAL datasets. We show qualitative illustration for MaskCLIP with custom categories.

1 Introduction

Panoptic segmentation [16] or image parsing [30] integrates the task of semantic segmentation [29] for background regions (e.g. “stuff” like “road”, “sky”) and instance segmentation [9] for foreground objects (e.g. “things” such as “person”, “table”). Existing panoptic segmentation methods [16, 15, 21, 33, 17] and instance segmentation approaches [9] deal with a fixed set of category definitions, which are essentially represented by categorical labels without semantic relations. DEtection TRansformer (DETR) [3] is a pioneering work that builds a Transformer-based architecture for object detection and panoptic segmentation.

Despite the recent development in open-vocabulary object detection [35, 18, 34] and open-vocabulary semantic labeling/segmentation [19, 7], the problem of open-vocabulary instance/panoptic segmentation has not been previously studied. A successful system for open-vocabulary panoptic segmentation will help significantly expand the capability of the current image understanding systems by performing dense image/instance segmentation for categories at an unprecedented scale for thousands and even tens of thousands custom categories with text specification, e.g. “red toy car”, “person holding a blue bag”, and “road with zebra lines”.

In this paper, we take the advantage of the existence of the pre-trained CLIP model [27] consisting of a vision embedding model and a text embedding model that are mapped to the same space. We first build a baseline method for open-vocabulary panoptic segmentation using CLIP without...
training. We then develop a new algorithm, MaskCLIP, that is a Transformer-based approach using
mask queries with the ViT-based CLIP backbone to efficiently and effectively utilize pre-trained
dense/local CLIP features. MaskCLIP consists of a Relative Mask Attention (RMA) module that is
seamlessly integrated with a pre-trained CLIP; MaskCLIP is distinct and advantageous compared with
existing approaches in three aspects: 1) it avoids the time-consuming operation to crop image patches
and compute feature from an external CLIP image model, such as the “CLIP on cropped regions”
method in [8]; 2) MaskCLIP also avoids the challenging student-teacher distillation processes such
as OVR-CNN [35] and ViLD [8] that face the problem of having a limited number of teacher objects
to train; 3) MaskCLIP also learns to refine masks beyond simple pooling in e.g., OpenSeg [7].
MaskCLIP expands the scope of existing CLIP models to open-vocabulary panoptic segmentation by
demonstrating encouraging and competitive results for panoptic segmentation and semantic labeling.

The contributions of our work are listed as follows.
• We develop an efficient method, MaskCLIP, to effectively utilize dense/local features in pre-
trained vision transformer (ViT) CLIP image encoder by avoiding the expensive process of
extracting image regions for external CLIP feature recomputation. MaskCLIP also skips the
challenging student-teacher distillation process that is facing limited teacher objects to learn from.
• We devise a pipeline to perform open-vocabulary panoptic segmentation by first training a class-
agnostic mask proposal network and then using the output coarse masks for efficient dense image
feature extraction and mask refinement.
• We tackle a new computer vision task, open-vocabulary panoptic segmentation, that performs
semantic and instance segmentation by allowing arbitrary specification of text inputs including
category names or linguistic descriptions.

2 Related Work

Panoptic segmentation. The panoptic segmentation [16] (or image parsing [30]) task integrates
semantic labeling/segmentation [29] and instance segmentation [9]. Previous methods that performs
panoptic segmentation can be roughly split into two categories: 1) Use two heads which are semantic
head and instance head and fuse the results to get panoptic results [15, 33]. 2) Treat panoptic
segmentation the same as instance segmentation i.e. regarding stuff as things [3, 5, 4]. We adopt the
second way here. Existing panoptic segmentation methods [16, 15, 21, 33, 17] all perform training
and testing based on a fixed set of category labels. Exemplar-based open-set panoptic segmentation
work exists [12] but its problem formulation is different from the open-vocabulary setting. Previous
works mainly integrate the mask proposal and mask classification. MaskRCNN [9] mentions that
using a class-agnostic mask proposal network also works well on the traditional fully supervised
setting. Lots of previous works on instance/panoptic segmentation could be modified into class-
agnostic mask proposal networks by removing the class label supervision. Our pipeline adopts a
class-agnostic mask proposal network at first.

Open vocabulary. Open vocabulary setting is gaining increasing popularity lately as traditional fully
supervised setting cannot handle unseen classes during testing. While real world vision applications
like scene understanding, self driving and robotics are commonly required to predict unseen classes.
Previous open-vocabulary attempts have been primarily made for object detection. ViLD [8] trains a
student model to distill the knowledge of CLIP. RegionCLIP [36] finetunes the pretrained CLIP
model to match the image areas with corresponding texts. OV-DETR [34] uses CLIP as an external
model to obtain the query embedding from CLIP model. Recently there is also some work made
for open-vocabulary semantic segmentation. LSeg [19] trains a CNN model to compute per-pixel
image feature to match with a pre-trained language embedding. OpenSeg [7] initializes its backbone
using the ALIGN [13] model and outputs dense/local features by directly pooling the generated
class-agnostic masks with the global image features.

Open-vocabulary panoptic/instance segmentation: an emerging task. As stated above, there
have been previous attempts made for open-vocabulary semantic segmentation, but not for open-
vocabulary panoptic/instance segmentation. LSeg [19] aims at dense pixel-wise semantic labeling
(without the instance-level formulation) by combining different labeling categories across multiple
image segmentation datasets using learned language embedding, but the learned CNN image features
in LSeg [19] are not exposed to representations beyond the training labeling categories. OpenSeg
[7] is potentially applicable for instance/panoptic segmentation, but OpenSeg is formulated to be
trained on captions which lack instance-level information that is fundamental for panoptic segmentation. Moreover, the class-agnostic proposal network in OpenSeg [7] takes the features learned by the backbone which are also used for dense feature extraction, making the class proposal not purely agnostic. The direct image feature pooling strategy in OpenSeg [7] is potentially another limiting factor towards the open-vocabulary panoptic segmentation. Nevertheless, no results for open-vocabulary panoptic/instance segmentation are reported in [7]. Open-set panoptic segmentation [12] is an exemplar based approach that requires categories to be known in advance, which is narrower than the open-vocabulary setting where categories of interest can be freely specified in prediction. Existing open-vocabulary object detection approaches [35, 8, 34] perform bounding-box based detection, which, although highly related, have a large difference to the panoptic segmentation task that requires dense pixel-wise prediction.

**Vision-Language pretraining.** After its initial release, the CLIP model [27] that is learned from large-scale image-text paired captioning datasets has received a tremendous amount of attention. Some other similar vision-language models have also been proposed later e.g. ALIGN [13], GLIP [20]. Many algorithms have been developed lately [34, 31, 36, 24, 26, 28] trying knowledge distillation from the CLIP model to benefit the down-stream tasks one way or the other by leveraging the rich semantic language information paired in the images. Here, we directly adopt the backbone of CLIP image model to train for open-vocabulary panoptic/semantic segmentation.

**Prompt Tuning.** In natural language processing, prompt tuning is a well-known technique that transfers a pretrained model to a different task [32, 18, 11, 39]. It doesn’t require finetuning for the whole model with only the prompt to be tuned which is more efficient. Our method shares similar idea with prompt tuning by introducing Mask Class Tokens to the pre-trained image model but we don’t need to tune the prompt.

### 3 Method

![Diagram of the pipeline]

**Figure 1: Illustration of the pipeline.** Our pipeline contains two stages. The first stage is a class-agnostic mask proposal network and the second stage is the MaskCLIP built on the pretrained CLIP ViT model. All the weights of the CLIP ViT model during training are fixed. Arrows in orange denote weight sharing. The embeddings’ weights of Mask Class Tokens are shared by Class Tokens in the CLIP ViT model and are fixed. RMA represents Relative Mask Attention which is built based on the CLIP ViT attention layer. RMA contains all the weights from CLIP ViT attention layer which are all fixed during training. Additional weights are added in RMA for further mask information utilization and mask refinement. The demo image we use here is from ADE20K [37].

Our pipeline, shown in Figure 1 contains two stages. The first stage is a class-agnostic mask proposal network. The second stage is built on the CLIP [27] ViT architecture. It takes the images and the course masks from the first stage as the input and will output refined masks along with the corresponding dense image features for further classification using the text embeddings from the CLIP text encoder.
3.1 Mask Class Tokens

In order to obtain dense image features for the corresponding masks or bounding boxes for further recognition or distillation, an easy way to do this is simply masking or cropping the image and then sending the obtained image to the pretrained image encoder. This method has been widely used in several open vocabulary object detection methods [36, 8]. The problem is that it’s not computation efficient ($N$ masks/boxes will lead to $N$ images and they will be computed through the image encoder independently) and also loses the ability to see the global image context information which is very important for recognizing some objects and stuff. For masking, another problem is that masks are in different shapes and simply masking the image will cause the resulted image to have transparent background which usually doesn’t exist in real images that are used for training in large language-vision models e.g. CLIP.

To solve this, we propose Mask Class Tokens for efficient feature extraction from images without losing the global image context information. In the original CLIP ViT-based visual encoder framework, the input of the network is $N$ image tokens and 1 class token. The final output of the class token will be used for the relation computation with the text embeddings. Our newly introduced $M$ Mask Class Tokens will be alongside with the image tokens and the class token. The embeddings’ weights of the Mask Class Token are provided by the class token in the pretrained CLIP ViT model and are fixed. Each Mask Class Token will output a corresponding dense image feature similar to the class token which outputs the feature of the whole image. To achieve this, we design an attention mask as following

$$
\mathcal{M} = \begin{bmatrix}
\mathcal{F}_{M \times N}^{(N+1) \times (N+1)} & \mathcal{T}_{(N+1) \times M}^{(N+1) \times M} \\
\mathcal{M}_{M \times M} & \mathcal{F}_{M \times M}^{1} & \mathcal{T}_{M \times M}^{1}
\end{bmatrix}
$$

(1)

in which $M$ is the number of Mask Class Tokens, $N$ is the number of image tokens, $\mathcal{T}_{m \times n}$ is an $m \times n$ True matrix, $\mathcal{F}_{m \times n}$ is an $m \times n$ False matrix and $\mathcal{M}'$ is defined as following:

$$
\mathcal{M}'_{i,j} = \begin{cases}
\text{False} & \text{if mask}_i \text{ contains at least one pixel in patch}_j \\
\text{True} & \text{otherwise.}
\end{cases}
$$

(2)

where True means that this position is masked out i.e. not allowed to attend and False otherwise. In this way, each Mask Class Token will learn from the corresponding mask area of the images. The image tokens are also interacting with each other which means the global information won’t lose. And it’s also very efficient since we don’t need to do redundant computing for each mask or finetune the pretrained model. However, the mask information are not fully utilized and they cannot be refined either. But we will see in the experiments later that simply adopting Mask Class Tokens to the pretrained CLIP model without any finetuning will already serve as a competitive baseline.

3.2 Relative Mask Attention

![Figure 2: Relative Mask Attention.](image-url)
To further utilize the mask information and refine the coarse masks, we propose Relative Mask Attention mechanism in our transformer. Our key design principle is to try not to change the CLIP features directly as this would destroy the learned relationship between the image features and text features in the CLIP model. Therefore, we adopt a way to only change the attention matrix in the transformer to learn a better linear combination of the values in the attention layers according to the mask information. As in Figure 2, our proposed Relative Mask Attention Mechanism only changes the attention matrix and refines the masks.

\( M_r \) is defined in Eq. 5. \( A'_{M,-N} \) is defined in Eq. 3. \( f_r \) is a two-layer convolutional network that maps the attention matrix to a mask residual.

Similar to relative positional encoding, we use a relative attention mechanism here. Let \( D \) be the dimension of the token embedding, for each Mask Class Token \( T_{MC} \in \mathbb{R}^{M \times D} \) with a corresponding mask \( K_i \in \mathbb{R}^{H \times W} \) whose shape is the same as the image, we use a similar way as for the images to get mask patch tokens \( T_{MP} \in \mathbb{R}^{M \times N \times D} \) in the computation of the attention. In our attention matrix, the Mask Class Tokens attending image tokens part will then be as following:

\[
A'_{M,-N} = \frac{\sum_c (\phi_{Q_m}(T_{MP}) \odot \phi_{K_m}(T_{IM}))}{\sqrt{D}}
\]

where \( T_{IM} \in \mathbb{R}^{N \times D} \) is image tokens, \( T_{MC} \in \mathbb{R}^{M \times D} \) is Mask Class Tokens, \( T_{MP} \in \mathbb{R}^{M \times N \times D} \) is Mask Patch Tokens \( \phi_Q, \phi_K, \phi_{Q_m}, \phi_{K_m} \) are linear transformations, \( \odot \) is element-wise product and \( \sum_c (\cdot)_c \) is the sum of the embedding dimension. \( \phi_{K_m}(T_{IM}) \in \mathbb{R}^{N \times D} \) will first be broadcast to \( \mathbb{R}^{M \times N \times D} \) before doing element-wise production.

The attention will also in turn be used for the refinement of the masks. The vanilla attention can be seen as a relationship between each mask area and all the image patches. Thus we utilize this to help our coarse masks be more accurate. The updating process of the masks is as following:

\[
M_r = M_c + f_r(\phi_Q(T_{MC}) \odot \phi_K(T_{IM}))
\]

where \( M_c, M_r \) denotes the coarse mask and refined mask respectively, \( f_r \) is a learnable non-linear function that maps the attention matrix to a mask residual.

### 4 Experiments

In this part, we train our proposed MaskCLIP method using COCO\[22\] training data and test on other datasets under the open vocabulary setting. Due to the novel setting of the open vocabulary panoptic segmentation task, we also compare our performance on open vocabulary semantic segmentation with previous methods. Apart from the quantitative results, we also provide qualitative results to validate our method that has good ability to learn dense image features and can support user-specified arbitrary categories.

#### 4.1 Datasets

**COCO:** COCO \[22\] includes 133 classes where 80 classes are things and 53 classes are stuff or background. There are 118k training images and 5k validation images. In our experiments, we first train the class-agnostic mask proposal network on COCO training dataset using the annotations of panoptic masks. Then we train our models on COCO training images in a supervised manner.

**ADE20K:** ADE20K \[37, 38\] contains 20,210 images and annotations for training and 2000 images and annotations for validation. It serves both panoptic segmentation and semantic segmentation. The full version (A-847) \[37\] includes 847 classes and the short version (A-150) \[38\] includes 150 classes.
We use the validation set in ADE20K for testing without any training on this dataset in which case we can test our model’s capability of open vocabulary segmentation.

**PASCAL Context**: PASCAL Context [25] contains 10,103 per-pixel annotations for images of PASCAL VOC 2010 [6], where 4998 for training and 5105 for validation. The full version (P-459) includes 459 classes and the short version includes 59 classes. This dataset serves as another benchmark testing our model’s open vocabulary segmentation ability.

### 4.2 Implementation Details

**Class-Agnostic Mask Proposal Network.** In our first stage, we train a class-agnostic mask proposal network using MaskRCNN[9] and Mask2Former[4] on COCO training data. We remove the class supervision during training thus making them class agnostic. The experiment setting we use for MaskRCNN is R50-FPN-1x. The backbone we use in Mask2Former is ResNet-50. All the training setting follows the default in their models.

**CLIP Baseline.** We design our first baseline as shown in Figure 3 by directly using the class-agnostic mask proposal network from the first stage and the pretrained CLIP model. We mask the images according to the masks from the class-agnostic mask proposal network and send the masked images to the CLIP model to get classification scores. The pretrained CLIP model we use is ViT-L/14@336px and the text inputs we use are simply the category names defined by each dataset. Those two settings keep the same with the following two methods for fair comparison.

**MaskCLIP w/o RMA Baseline.** Our second baseline is based on the Mask Class Tokens which doesn’t use the Relative Mask Attention mechanism. Instead of masking the images and sending the resulted images directly to the CLIP model for feature extraction, we use Mask Class Tokens to acquire the corresponding dense image features. The obtained image features will then be used for further open vocabulary classification.

The two baselines above don’t need any training in the second stage and can be used to directly perform the open vocabulary tasks. We will demonstrate that the second baseline is better at feature extraction in both quantitative results and qualitative results under the open vocabulary setting and show the effectiveness and efficiency of the proposed Mask Class Tokens.

**MaskCLIP.** In our MaskCLIP method, we still use the CLIP ViT-L/14@336px pretrained model as with the previous two. This model has 24 attention layers and we add Relative Mask Attention in four of them which is 6, 12, 18, 24. We use AdamW[23] as our optimizer and the learning rate is set to 0.0001. We train our model on COCO training data for 10k iterations with a batchsize of 8. The training takes around 3h on 8 Nvidia A5000 GPUs.

In next three parts, we evaluate our methods on open vocabulary panoptic, instance segmentation, and semantic segmentation tasks. The class-agnostic mask proposal networks we use in those methods are trained using Mask2Former other than noted.

### 4.3 Open-Vocabulary Panoptic Segmentation

First, we compare our MaskCLIP with the two baselines on ADE20K validation set under the open vocabulary panoptic segmentation setting. The results are presented in Table 1. As can be seen from the table, the MaskCLIP w/o RMA baseline performs better on all the metrics in panoptic segmentation setting which demonstrates that our feature extraction method is better than the vanilla way in this setting. It extracts the features without the need to changing the input and can simultaneously extract multiple mask area features easily. For 100 masks' feature extraction in a single image, the CLIP baseline takes ~3s on a single 3090 GPU while the MaskCLIP w/o RMA baseline only takes ~0.6s which is ~4x faster. Our MaskCLIP beats both baselines significantly as it utilizes the accurate mask information and refines the masks during the feature extraction process.
Table 1: **Results on open-vocabulary panoptic segmentation using the ADE20k validation dataset.** PQ, SQ, RQ are three metrics for evaluating the panoptic segmentation qualities. th and st represent thing and stuff classes respectively. RMA refers to the Relative Mask Attention (RMA) module shown in Figure 1. ↑ indicates the higher the better.

| Method                        | PQ ↑ | PQ*↑ | PQ**↑ | SQ ↑ | SQ*↑ | SQ**↑ | RQ ↑  | RQ*↑  | RQ**↑ |
|-------------------------------|------|------|-------|------|------|-------|-------|-------|-------|
| CLIP Baseline                 | 8.207| 8.473| 7.675 | 53.124| 52.661| 54.048| 10.534| 10.883| 9.835 |
| MaskCLIP w/o RMA              | 9.565| 9.922| 10.852| 62.507| 62.268| 62.985| 12.645| 11.758| 14.418|
| MaskCLIP (MaskRCNN)           | 12.860| 11.242| 16.095| 64.008| 64.183| 63.658| 16.803| 14.968| 20.473|
| MaskCLIP                      | 15.121| 13.536| 18.290| 70.479| 70.021| 71.396| 19.211| 17.448| 22.737|

Figure 4: **Qualitative results on ADE20K panoptic segmentation.** The images are taken from the ADE20K validation set. We use the class names directly from the ADE20K 150 classes as the text inputs. Three images are presented here using our MaskCLIP model along with the two baselines.
In this part, we show two sets of images to demonstrate our model capability. The first is the qualitative results on ADE20K. We compare our method with the two baselines in Figure 4. It can be seen that our method performs much better than the two baselines. The results from the first column show that due to the lack of global information, CLIP baseline fails to predict the floor. Instead, it predicts skyscraper. This is an easy case but if only floor area is provided, it does have some similarities with the wall of a skyscraper. While the MaskCLIP w/o RMA baseline and MaskCLIP model can predict the floor correctly with the global image context information.

The second set of images we're presenting are in Figure 5. These figures show our capability of specifying any arbitrary classes in performing panoptic segmentation task. The results show that though we train a new model based on the CLIP model without any distillation methods, we can still preserve the CLIP image features very well. Our model doesn’t have a clear bias towards the base classes in the training set and could tell the difference very well that have no chance to learn in the COCO training: e.g. toy vs real and filled vs empty.

![Images](image1.png)

Figure 5: User-specified class panoptic segmentation. The labels above are the text inputs we used for testing the images. Texts in bold are novel classes i.e. don’t exist in the labels of COCO training data. (a) Our model is able to distinguish object properties of real rabbit and toy rabbit. (b) This example shows that our model is potential for fine-grained classifications and does not have bias toward the base classes. (c) Our results show that it can tell the difference between the filled status and empty status of bottles.

### 4.4 Open-Vocabulary Semantic Segmentation/Labeling

Table 2: Results on open-vocabulary semantic segmentation. A-150 and A-847 represent the ADE20K dataset with 150 classes and 847 classes respectively. P-459 and P-59 represents the PASCAL Context dataset with 459 classes and 59 classes respectively. All the results use the mIoU metric. Results of ALIGN, ALIGN w/ proposals, LSeg+ and OpenSeg are cited from [7]. All the methods presented here don’t use extra data other than COCO for training. RMA refers to the Relative Mask Attention (RMA) module shown in Figure 1.

| Method              | COCO Training Data | A-150 ↑ | A-847 ↑ | P-459 ↑ | P-59 ↑ |
|---------------------|--------------------|---------|---------|---------|--------|
| ALIGN [13]          | None               | 10.7    | 4.1     | 3.7     | 15.7   |
| ALIGN w/ proposals  [13] | Masks              | 12.9    | 5.8     | 4.8     | 22.4   |
| LSeg+ [19]          | Masks + Labels     | 18.0    | 3.8     | 7.8     | 46.5   |
| OpenSeg [7]         | Masks + Captions   | 21.1    | 6.3     | 9.0     | 42.1   |
| CLIP Baseline       | Masks              | 13.8    | 5.2     | 5.2     | 25.3   |
| MaskCLIP w/o RMA    | Masks              | 14.9    | 5.6     | 5.3     | 26.1   |
| MaskCLIP (MaskRCNN) | Masks + Labels     | 22.4    | 6.8     | 9.1     | 41.3   |
| MaskCLIP            | Masks + Labels     | 23.7    | 8.2     | 10.0    | 45.9   |

We also use our method to compare with open-vocabulary semantic segmentation as in Table 2. The setting is similar, they all train on COCO panoptic training set and test on ADE20K validation set.
On the four datasets we test, MaskCLIP reaches the state-of-the-art results on three of them with only P-59 slightly lower.

To compare with previous methods, we also provide a semantic segmentation comparison in Figure 6. Results on ALIGN++ and OpenSeg are directly from [7] and we run the same image using our MaskCLIP model. It can be seen that due to the open vocabulary setting, some similar classes may be mistakenly classified e.g. all three methods predict the house in this image while the ground truth is building.

4.5 Open-Vocabulary Instance Segmentation

In this part, we present the results on open vocabulary instance segmentation in Table 3. Since instance segmentation can be regarded as “thing-only” panoptic segmentation, we directly apply our model trained on coco panoptic dataset to the instance segmentation task. MaskCLIP with different class-agnostic mask proposal networks perform better than CLIP Baseline and MaskCLIP w/o RMA in general.

Table 3: Results on open-vocabulary instance segmentation using the ADE20k validation dataset. RMA refers to the Relative Mask Attention (RMA) module shown in Figure 1. ↑ indicates the higher the better.

| Method            | AP ↑ | AP50↑ | AP75↑ | APs ↑ | APM↑ | APL↑ |
|-------------------|------|-------|-------|-------|------|------|
| CLIP Baseline     | 3.974| 6.090 | 4.288 | 1.057 | 4.891| 10.881|
| MaskCLIP w/o RMA  | 4.263| 6.696 | 4.402 | 1.568 | 4.867 | 10.943|
| MaskCLIP (MaskRCNN) | **6.164** | **12.072** | **5.775** | **1.039** | **7.010** | **16.337** |
| MaskCLIP          | 5.989| 9.739 | 6.209 | 1.569 | 7.369 | 15.161|

We also show qualitative results on ADE20K comparing the CLIP Baseline, MaskCLIP w/o RMA and MaskCLIP. As shown in Figure 7 MaskCLIP is much better than the CLIP Baseline and MaskCLIP w/o RMA. Since the mask proposal network is class-agnostic, some masks that are not objects would actually be predicted in which case the classification later will be very important as it may be classified as some object classes in the dataset. The visualization results of MaskCLIP contains fewer non-object masks and are more accurate in class prediction.

Figure 7: Qualitative results on ADE20K instance segmentation.
5 Ablation Study

5.1 Mask Refinement

In our Relative Mask Attention part, the attention layer will use the accurate mask information to learn a better attention matrix and the mask will also use the attention information to gradually refine itself. In this ablation study, we only let the attention matrix learn from the mask without any mask refinement. And we get the results in Table 4. Since the SQ reflects the segmentation quality, we care more about SQ here. It can be seen that MaskCLIP performs slightly better than that without the mask refinement which demonstrates the effectiveness of the mask refinement.

Table 4: Ablation Study on Mask Refinement. Results on ADE20K validation set are reported here. Both methods are trained on COCO and tested on ADE20K validation dataset.

|                | PQ↑ | PQTh↑ | PQSt↑ | SQ↑ | SQTh↑ | SQSt↑ |
|----------------|-----|-------|-------|-----|-------|-------|
| MaskCLIP w/o mask refinement | 13.624 | 13.253 | 14.368 | 66.361 | 67.715 | 63.653 |
| MaskCLIP       | 15.121 | 13.536 | 18.290 | 70.479 | 70.021 | 71.396 |

5.2 Relative Mask Attention in Different Layers

In this part we conduct an ablation study on using different layers for relative mask attention. Since our pretrained CLIP model is fixed during the whole training procedure, whether each layer would help the final results remains a question. We use four different kinds of layers combination of the layers in this part and provide the results in Table 5. We can see that the last layer is a key part of our results since the features are gradually learned through all the attention layer. Though the last four layers’ features should be the best, the performance wouldn’t be better if Relative Mask Attention is only used in the last four layers. This is also reasonable since the network should not have the accurate mask information too late.

Table 5: Ablation Study on Relative Mask Attention Layers in different layers. All the methods are trained on COCO and tested on ADE20K validation dataset. The pretrained CLIP ViT-L/14@336px model has 24 layers and we replace four of them with our relative mask attention to fully utilize the accurate mask information and refine the masks.

| Different Layers | PQ  | PQTh | PQSt |
|------------------|-----|------|------|
| 1, 7, 13, 19     | 11.241 | 10.519 | 12.686 |
| 3, 9, 15, 21     | 11.372 | 10.141 | 13.835 |
| 21, 22, 23, 24   | 14.673 | 14.048 | 15.922 |
| 6, 12, 18, 24    | 15.121 | 13.536 | 18.290 |

6 Conclusion

Limitation: Since our work builds on top of the pretrained CLIP [27] models, the overall performance of our system is largely decided by the quality of the CLIP models. Although our MaskCLIP method shows encouraging results for the open-vocabulary panoptic/semantic/instance segmentation tasks, there is nerveless still a large room for improvement in the performance.

In this paper, we have presented a new algorithm, MaskCLIP, to tackle an emerging computer vision task, open-vocabulary panoptic segmentation. MaskCLIP is a Transformer-based approach using mask queries with the ViT-based CLIP backbone to efficiently and effectively utilize pre-trained dense/local CLIP features. MaskCLIP consists of a Relative Mask Attention (RMA) module that is seamlessly integrated with a pretrained CLIP. MaskCLIP is distinct compared with existing approaches in open-vocabulary semantic segmentation/object detection by building an integrated encoder module for segmentation mask refinement and image feature extraction with a pre-trained CLIP image model. Encouraging experimental results on open-vocabulary panoptic segmentation, open-vocabulary semantic segmentation, and open-vocabulary instance segmentation have been obtained.

7 Acknowledgement

This work is supported by NSF Award IIS-2127544. We thank Xiang Zhang and Boyi Li for helpful discussions.
References

[1] Abhijit Bendale and Terrance Boult. Towards open world recognition. In CVPR, 2015.

[2] Zhaowei Cai, Gukyeong Kwon, Avinash Ravichandran, Erhan Bas, Zhuowen Tu, Rahul Bhotika, and Stefano Soatto. X-detr: A versatile architecture for instance-wise vision-language tasks. arXiv preprint arXiv:2204.05626, 2022.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In ECCV, pages 213–229, 2020.

[4] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. arXiv preprint arXiv:2112.01527, 2021.

[5] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems, 34, 2021.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html.

[7] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Open-vocabulary image segmentation. arXiv preprint arXiv:2112.12143, 2021.

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

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In ICCV, pages 2961–2969, 2017.

[10] Ryota Hinami and Shin’ichi Satoh. Discriminative learning of open-vocabulary object retrieval and localization by negative phrase augmentation. arXiv preprint arXiv:1711.09509, 2017.

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

[12] Jaedong Hwang, Seoung Wug Oh, Joon-Young Lee, and Bohyung Han. Exemplar-based open-set panoptic segmentation network. In CVPR, pages 1175–1184, 2021.

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning, pages 4904–4916, 2021.

[14] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In CVPR, pages 1780–1790, 2021.

[15] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In CVPR, 2019.

[16] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In CVPR, 2019.

[17] Justin Lazarow, Kwonjoon Lee, Kunyu Shi, and Zhuowen Tu. Learning instance occlusion for panoptic segmentation. In CVPR, 2020.

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

[19] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In ICLR, 2022.

[20] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. arXiv preprint arXiv:2112.03857, 2021.

[21] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. In CVPR, 2019.
Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV, pages 740–755, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. arXiv preprint arXiv:2104.08860, 2021.

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. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In CVPR, pages 2085–2094, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, pages 8748–8763, 2021.

Zhuowen Tu. Auto-context and its application to high-level vision tasks. In CVPR, 2008.

Zhuowen Tu, Xiangrong Chen, Alan L Yuille, and Song-Chun Zhu. Image parsing: Unifying segmentation, detection, and recognition. International Journal of computer vision, 63(2):113–140, 2005.

Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In CVPR, 2019.

Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. arXiv:2203.11876, 2022.