Extract Free Dense Labels from CLIP

Chong Zhou[1], Chen Change Loy[1], and Bo Dai[2⋆]

[1] S-Lab, Nanyang Technological University
[2] Shanghai AI Laboratory
{chong033, ccloy}@ntu.edu.sg daibo@pjlab.org.cn

Abstract. Contrastive Language-Image Pre-training (CLIP) has made a remarkable breakthrough in open-vocabulary zero-shot image recognition. Many recent studies leverage the pre-trained CLIP models for image-level classification and manipulation. In this paper, we wish examine the intrinsic potential of CLIP for pixel-level dense prediction, specifically in semantic segmentation. To this end, with minimal modification, we show that MaskCLIP yields compelling segmentation results on open concepts across various datasets *in the absence of annotations and fine-tuning*. By adding pseudo labeling and self-training, MaskCLIP+ surpasses SOTA transductive zero-shot semantic segmentation methods by large margins, *e.g.*, mIoUs of unseen classes on PASCAL VOC/PASCAL Context/COCO Stuff are improved from 35.6/20.7/30.3 to 86.1/66.7/54.7. We also test the robustness of MaskCLIP under input corruption and evaluate its capability in discriminating fine-grained objects and novel concepts. Our finding suggests that MaskCLIP can serve as a new reliable source of supervision for dense prediction tasks to achieve annotation-free segmentation. Source code is available here.

1 Introduction

Large-scale visual-language pre-training models such as CLIP [45] capture expressive visual and language features. Various downstream vision tasks, *e.g.*, text-driven image manipulation [42], image captioning [25], view synthesis [30], and object detection [19], have attempted to exploit such features for improved generality and robustness. For instance, conducting zero-shot image classification based on raw CLIP features leads to a competitive approach that matches the performance of fully-supervised counterparts [45].

In this paper, we take a step further to explore the applicability of CLIP features for pixel-level dense prediction tasks such as semantic segmentation. This investigation is meaningful in that previous studies mainly leverage CLIP features as a global image representation. In contrast, our exploration wishes to ascertain the extent of CLIP features in encapsulating object-level and local semantics for dense prediction. Different from the conventional pre-training task of image classification on iconic images, CLIP learns from images of complex scenes and their descriptions in natural language, which (1) encourages it

⋆ Bo Dai completed this work when he was with S-Lab, NTU.

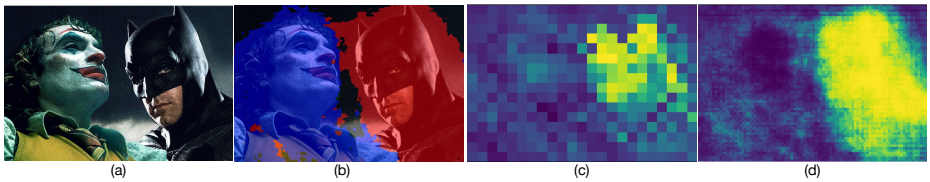

Fig. 1: Here we show the original image in (a), the segmentation result of MaskCLIP+ in (b), and the confidence maps of MaskCLIP and MaskCLIP+ for *Batman* in (c) and (d) respectively. Through the adaptation of CLIP, MaskCLIP can be directly used for segmentation of fine-grained and novel concepts (e.g., *Batman* and *Joker*) without any training operations and annotations. Combined with pseudo labeling and self-training, MaskCLIP+ further improves the segmentation result.

to embed local image semantics in its features, (2) enables it to learn concepts in open vocabulary, and (3) captures rich contextual information, such as the co-occurrence/relation of certain objects and priors of the spatial locations. We believe all these merits contribute significantly to its potential for dense prediction tasks.

In this paper, we summarize both our success and failure experience on leveraging CLIP features for dense prediction. We find it essential to not break the visual-language association in the original CLIP feature space. In our earlier exploration, we experienced failures with the attempt to fine-tune the image encoder of CLIP for the segmentation task, *e.g.*, initializing DeepLab [5] with the weights of CLIP's image encoder and fine-tune the backbone on segmentation. In addition, we found it is of utmost importance to avoid any unnecessary attempts to manipulate the text embeddings of CLIP. Such an approach would fail in segmenting unseen classes.

In our successful model, named **MaskCLIP**, we show that one can simply extract dense patch-level features from the CLIP's image encoder, *i.e.*, the *value* features of the last attention layer, without breaking the visual-language association. Classification weights for dense prediction, which are essentially 1×1 convolutions, can be directly obtained from the text embeddings of CLIP's text encoder without any deliberate mapping. In our empirical study, MaskCLIP yields reasonable predictions in both quantitative performance measured by mIoU metric and qualitative results. Besides, MaskCLIP can be based on all variants of CLIP, including ResNets and ViTs. And we provide side-by-side comparisons between the two popular backbone networks. We also propose two mask refinement techniques for MaskCLIP to further improve its performance, namely *key smoothing* and *prompt denoising*, both require no training. Specifically, key smoothing computes the similarity between the *key* features (of the last attention layer) of different patches, which are used to smooth the predictions. Prompt denoising removes prompts with classes that unlikely exist in the image, thus with fewer distractors, predictions become more accurate.

However, it is hard to further improve the segmentation capacity of MaskCLIP as its architecture is restricted to be the image encoder of CLIP. To relax MaskCLIP from the architectural constraint and to incorporate more advanced architectures such as PSPNet [55] and DeepLab [5,6], we notice that instead of deploying MaskCLIP at the inference time, we can deploy it at the training time, where it serves as a generalizable and robust annotator that provides high-quality pseudo labels. Together with a standard self-training strategy, the resulting model, termed **MaskCLIP+**, achieves a strikingly remarkable performance.

Apart from annotation-free and open-vocabulary segmentation, MaskCLIP+ can also be applied to the transductive zero-shot semantic segmentation task, where MaskCLIP only generates pseudo labels for the unseen classes. On the three standard segmentation benchmarks, namely PASCAL VOC [15], PASCAL Context [38], and COCO Stuff [2], MaskCLIP+ improves the state-of-the-art results in terms of mIoU of unseen classes, by 50.5%, 46%, and 24.4%, respectively ($35.6 \rightarrow 86.1$, $20.7 \rightarrow 66.7$, and $30.3 \rightarrow 54.7$). Thanks to the the generality and robustness of CLIP features, MaskCLIP+ can be readily applied to various extended settings of semantic segmentation, including the segmentation of fine-grained classes (*e.g.*, attribute-conditioned classes like *white car* and *red bus*) or novel concepts (such as *Batman* and *Joker* as shown in Figure 1), as well as the segmentation of moderately corrupted inputs. We show more interesting results in the experiment section.

Semantic segmentation is notorious for its high dependency on labeled training data. Many methods have been explored to get around such stringent requirement, *e.g.*, through using weak labels like image tags, bounding boxes, and scribbles. Our study, for the first time, shows that features learned via large-scale visual-language pre-training can be readily used to facilitate open vocabulary dense prediction. The proposed model, MaskCLIP, shows promising potential in providing rich and meaningful dense pseudo labels for training existing methods.

2 Related Work

Transferable Representation Learning. Pre-training is widely used for dense prediction tasks. Yosinski *et al*. [52] show that ImageNet [11] pre-training greatly speeds up the convergence of the downstream object detection task. Later, extensive research is conducted on making the pre-training a human-labor-free process. In particular, self-supervised representation learning constructs pretext tasks [14,13,39] or relies on contrastive learning [22,7,10], clustering[3], and bootstrapping[18] to obtain supervision signals. Another line of work seeks to learn visual representation from natural language. Some studies [44,16,17,48,12,53] propose to learn from image-caption pairs. Recently, CLIP [45] and ALIGN [31] perform contrastive learning on very large-scale web-curated image-text pairs and show promising pre-trained representations with impressive zero-shot transferability. The success of CLIP inspires a new way of studies that transfer the pre-trained CLIP model to various downstream tasks such as text-driven im-

age manipulation [42], image captioning [25], view synthesis [30], and object detection [19]. Different from these methods that typically apply CLIP right off the shelf for image encoding, we explore ways to adapt CLIP for pixel-level dense prediction. A concurrent work, DenseCLIP [46], aims to better fine-tune the CLIP pre-trained weights on target semantic segmentation datasets without keeping the zero-shot transferability, which are different from our setting. To examine the intrinsic potential of CLIP for dense prediction tasks, we refrain from any fine-tuning and major architectural modification.

Zero-Shot Visual Recognition. Zero-shot learning aims at classifying instances of those categories that are not seen during training. Common clues to infer unseen categories include shared attributes and visual-semantic mapping. As the latter does not require extra annotations, the paradigm is well-suited for zero-shot dense prediction tasks. Zhao *et al*. [54] project image pixel features and word concepts into a joint space. Kato *et al*. [32] fuse semantic features into visual features as guidance. ZS3Net [1] proposes to generate fake pixel-level features from semantic features for the unseen. SPNet [50] learns a projection from visual space to semantic space. Other studies like [26], [34], and [20], improve the generative ZS3Net in terms of uncertainty, structural consistency, and context, respectively, while STRICT [41] boosts the SPNet through self-training. Depending on whether the unlabeled pixels are observed during training, the setting can be split into inductive (not observed) and transductive. We show that the proposed MaskCLIP not only achieves new SOTA on the zero-shot segmentation setting but can also deal with more difficult settings where all the categories are unseen during training.

Self-Training. Self-training leverages the model trained on labeled data to generate pseudo labels for the unlabeled, which then are used to iteratively improve the previous model. Self-training has firstly become popular in the semi-supervised classification task [33,29,35,43] and is also recently applied in the semi-supervised/zero-shot semantic segmentation settings [27,37,4,28,40,36,56,8]. Our MaskCLIP+ adopts the same philosophy where the pseudo labels are obtained from both frozen MaskCLIP and MaskCLIP+ itself.

3 Methodology

Our study serves as an early attempt that explores the applicability of CLIP features for pixel-level dense prediction tasks. We start with a brief introduction of CLIP and a naïves solution as the preliminary, followed by presenting the proposed MaskCLIP in detail.

3.1 Preliminary on CLIP

CLIP [45] is a visual-language pre-training method that learns both visual and language representations from large-scale raw web-curated image-text pairs. It consists of an image encoder $\mathcal{V}(\cdot)$ and a text encoder $\mathcal{T}(\cdot)$, both jointly trained to respectively map the input image and text into a unified representation space.

CLIP adopts contrastive learning as its training objective, where ground-truth image-text pairs are regarded as positive samples, and mismatched image-text pairs are constructed as negative ones. In practice, the text encoder is implemented as a Transformer [49]. As for the image encoder, CLIP provides two alternative implementations, namely a Transformer and a ResNet [23] with global attention pooling layer. Our method can be based on both encoder architectures.

We believe CLIP has inherently embedded local image semantics in its features as it learns to associate image content with natural language descriptions, the latter of which contain complex and dense semantic guidance across multiple granularities. For example, to correctly identify the image corresponds to the description *the man at bat readies to swing at the patch while the umpire looks on* [9], CLIP must divide image semantics into local segments and properly align image semantics with singular mentioned concepts like *man, bat, swing, patch, man at bat, man at patch*, and *man readies to swing*, instead of handling the image as a whole. Such uniqueness is absent from training with solely image labels.

3.2 Conventional Fine-Tuning Hinders Zero-Shot Ability

The current de facto pipeline of training a segmentation network is (1) initializing the backbone network with the ImageNet [11] pre-trained weights, (2) adding segmentation-specific network modules with randomly initialized weights, and (3) jointly fine-tuning the backbone and newly added modules.

It is natural to follow these standard steps to adapt CLIP for segmentation. Here, we start our exploration by applying this pipeline on DeepLab [5] with two CLIP-specific modifications. Specifically, we first replace the ImageNet pre-trained weights with weights of the image encoder of CLIP. Second, we adopt a mapper $\mathcal{M}$ that maps text embeddings of CLIP to the weights of DeepLab classifier (the last 1×1 convolutional layer). The modified model can be formulated as follows:

$$\text{DeepLab}(x) = \mathcal{C}_\phi(\mathcal{H}(\mathcal{V}_{*l}(x))), \tag{1}$$

$$\phi = \mathcal{M}(t), \tag{2}$$

where $\mathcal{V}_{*l}(\cdot)$ denotes the DeepLab backbone, which is a ResNet dilated by a factor of l. $H(\cdot)$ denotes the randomly initialized ASPP module [5], and $\mathcal{C}_\phi(\cdot)$ is the DeepLab classifier, whose weights, denoted as ϕ, are determined by the text embedding of CLIP via the mapper $\mathcal{M}$. Ideally, by updating the classifier weights with the corresponding text embedding, the adapted DeepLab is able to segment different classes without re-training.

To evaluate the segmentation performance of this modified DeepLab on both seen and unseen classes, we train it on a subset of classes in the dataset, considering the remaining classes as unseen ones. We have tried a series of mapper architectures. Although they perform well on seen classes, in all these cases the modified DeepLab fails to segment unseen classes with satisfying performance.

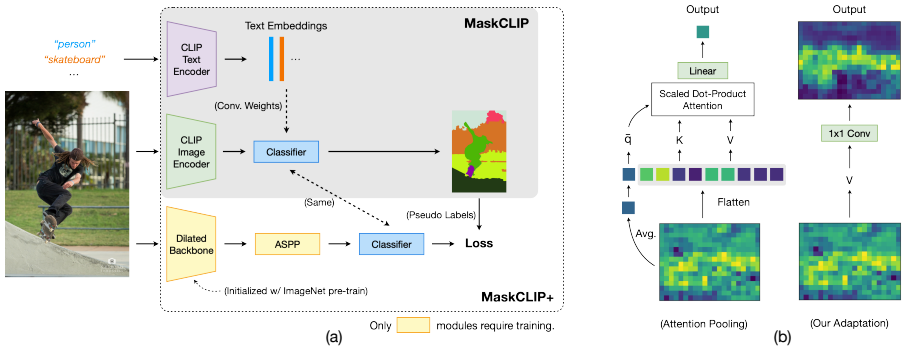

Fig. 2: **Overview of MaskCLIP/MaskCLIP+.** Compared to the conventional fine-tuning method, the key to the success of MaskCLIP is keeping the pre-trained weights frozen and making minimal adaptation to preserve the visual-language association. Besides, to compensate for the weakness of using the CLIP image encoder for segmentation, which is designed for classification, MaskCLIP+ uses the outputs of MaskCLIP as pseudo labels and trains a more advanced segmentation network such as DeepLabv2 [5]

We hypothesize that this is mainly because the original visual-language association of CLIP features has been broken: (1) the backbone is slightly different from the image encoder in terms of network architecture; (2) weights initialized from the image encoder have been updated during fine-tuning; (3) an extra mapper, which is trained only on data of seen classes, is introduced therefore leading to insufficient generality.

3.3 MaskCLIP

Failing the fine-tuning attempt, we turn to a solution that avoids introducing additional parameters and modifying the feature space of CLIP. To this end, we carefully revisit the image encoder of CLIP, especially its unique global attention pooling layer. As shown in Figure 2(b), different from conventional global averaged pooling, the image encoder of CLIP adopts a Transformer-style multi-head attention layer where globally average-pooled feature works as the query, and feature at each spatial location generates a key-value pair. Consequently, the output of this layer is a spatial weighted-sum of the incoming feature map followed by a linear layer $\mathcal{F}(\cdot)$:

$$\begin{aligned}\text{AttnPool}(\bar{q}, k, v) &= \mathcal{F}(\sum_i \text{softmax}(\frac{\bar{q}k_i^\intercal}{C})v_i) \\ &= \sum_i \text{softmax}(\frac{\bar{q}k_i^\intercal}{C})\mathcal{F}(v_i), \end{aligned} \tag{3}$$

$$\bar{q} = \text{Emb}_\text{q}(\bar{x}),\ k_i = \text{Emb}_\text{k}(x_i),\ v_i = \text{Emb}_\text{v}(x_i), \tag{4}$$

where C is a constant scaling factor and $\text{Emb}(\cdot)$ denotes a linear embedding layer[3]. x_i represents the input feature at spatial location i and $\bar{x}$ is the average of all x_i. The outputs of the Transformer layer serve as a comprehensive representation of the whole image. We believe that this is possible because $\mathcal{F}(v_i)$ computed at each spatial location already captures a rich response of local semantics that correspond well with tokens in the text embeddings of CLIP.

Based on such a hypothesis, as shown in Figure 2(b), we directly modify the image encoder of CLIP in our new attempt: (1) removing the query and key embedding layers; (2) reformulating the value-embedding layer and the last linear layer into two respective 1×1 convolutional layers. Moreover, we keep the text encoder unchanged and it takes prompts with target classes as the input. The resulting text embedding of each class is used as the classifier. We name the resulting model as MaskCLIP since it yields pixel-level mask predictions instead of a global image-level prediction. We then evaluate MaskCLIP on various standard segmentation benchmarks as well as web-crawled images. As shown in Figure 1, MaskCLIP can output reasonable results without any fine-tuning nor annotations. More qualitative results and quantitative results with respect to the mIoU metric are included in the experiment section.

One might argue that, since the global attention pooling is a self-attention layer, even without modification, it can also generate dense features. However, since query $\bar{q}$ is the only query trained during the CLIP pre-training, this naïves solution fails. We treat this solution as the baseline and compare its results with ours in the experiments. Moreover, the Transformer layer in ViT is very similar to the global attention pooling. In fact, the only two differences are: (1) the global query is generated by a special [CLS] token instead of the average among all spatial locations; (2) Transformer layer has a residual connection. Therefore, by replacing $\bar{q}$ with $q_{[cls]}$ and adding input x to the output, MaskCLIP can work with the ViT backbone.

Despite the simplicity of MaskCLIP in comparison to existing segmentation approaches, the proposed method enjoys multiple unique merits inherited from CLIP. First, MaskCLIP can be used as a free segmentation annotator to provide rich and novel supervision signals for segmentation methods working with limited labels. Second, since the visual-language association of CLIP is retained in MaskCLIP, it naturally possesses the ability to segment open vocabulary classes, as well as fine-grained classes described by free-form phrases, such as *white car* and *red bus*. Third, since the CLIP is trained on raw web-curated images, CLIP demonstrates great robustness to natural distribution shift [45] and input corruptions [47]. We verify that MaskCLIP preserves such robustness to some extent.

Key Smoothing and Prompt Denoising. To further improve the performance of MaskCLIP, we propose two refinement strategies, namely *key smoothing* and *prompt denoising*. Recall that, in Eq. 3, besides $\bar{q}$, key features k_i also get trained during CLIP pre-training. However, in the original MaskCLIP, k_i is

[3] Here we have simplified the formula by ignoring the channel-wise splitting and concatenation.

simply discard. Hence, here we seek to utilize this information to refine the final output. Key features can be viewed as the descriptor of the corresponding patch, therefore patches with similar key features should yield similar predictions. With this hypothesis, we propose to smooth the predictions with:

$$\text{pred}_i = \sum_j \cos(\frac{k_i}{\|k_i\|_2}, \frac{k_j}{\|k_j\|_2})\text{pred}_i, \tag{5}$$

where k_i and pred_i are key features and class confidence predictions at spatial location i, while $\|\cdot\|_2$ and $\cos(\cdot)$ denote L2 normalization and cosine similarity. We name this strategy as key smoothing.

In addition, we also observe that when dealing with many target classes, since only a small proportion of the classes appear in a single image, the rest classes are in fact distractors and undermine the performance. Therefore, we propose prompt denoising, which removes the prompt with target class if its class confidence at all spatial locations is all less than a threshold $t = 0.5$.

3.4 MaskCLIP+

While MaskCLIP does not require any training, its network architecture is rigid because it adopts the image encoder of CLIP. To relax it from this constraint and benefit from more advanced architectures tailored for segmentation, such as DeepLab [5] and PSPNet [55], we propose MaskCLIP+. Instead of directly applying MaskCLIP for test-time prediction, MaskCLIP+ regard its predictions as training-time pseudo ground-truth labels. Together with an adopted self-training strategy, MaskCLIP+ is thus free from the restriction on its backbone architecture. As shown in Figure 2(a), we take DeepLabv2 [5] as the backbone of MaskCLIP+ to ensure a fair comparison with previous segmentation methods.

MaskCLIP-Guided Learning. In MaskCLIP+, we leverage the predictions of MaskCLIP to guide the training of another target network comprising an architecture tailored to segmentation task. In parallel to the target network, we feed the same pre-processed image input to the MaskCLIP and use the predictions of MaskCLIP as pseudo ground-truth labels to train the target network. In addition, we replace the classifier of the target network with that of MaskCLIP, to preserve the network's ability for open vocabulary prediction.

MaskCLIP-guided learning is also applicable in the transductive zero-shot segmentation setting. Specifically, while pixels of both seen and unseen classes are observed, only annotations of seen classes are available. In this case, we only use MaskCLIP to generate pseudo labels for the unlabeled pixels. Compared to SOTA methods, MaskCLIP+ obtains remarkably better results across three standard benchmarks, namely PASCAL VOC 2012 [15], PASCAL Context [38], and COCO Stuff [2], where the results of MaskCLIP+ are even on par with that of fully-supervised baselines.

We note that some related attempts [19,51], targeting object detection, perform knowledge distillation between the image-level visual features of CLIP and the features of a target model. Different from such feature-level guidance, we

adopt pseudo labels in our case. This is because our target network, with a segmentation-tailored architecture, is structurally distinct from the image encoder of CLIP. Therefore, distillation by feature matching may be a sub-optimal strategy. In fact, as reported by [19], under zero-shot setting, such feature-level guidance indeed results in conflicts between the performance of seen and unseen classes. On the contrary, by adopting pseudo labels in MaskCLIP+, we do not observe any performance drop on seen classes.

Self-Training. It is expected that after certain training iterations, the target network guided by MaskCLIP will outperform MaskCLIP, rendering the latter suboptimal for further guidance as it gradually becomes an inferior model over time. Empirically, we also find that MaskCLIP-guided learning reaches an upper bound at around 1/10 of the standard training schedule. To further improve the performance, we swap out MaskCLIP and let the target model generate pseudo labels for itself. This is commonly referred to as self-training.

4 Experiments

Datasets. We conduct experiments on three standard segmentation benchmarks, namely PASCAL VOC 2012 [15], PASCAL Context [38], and COCO Stuff [2]. PASCAL VOC 2012 contains 1,426 training images with 20 object classes plus a background class. Following common practice, we augment it with the Semantic Boundaries Dataset [21]. PASCAL Context labels PASCAL VOC 2010 (4,998/5,105 train/validation images) with segmentation annotations of 520 object/stuff classes, from which the most common 59 classes are treated as foreground while the rest are regarded as background. COCO Stuff extends the COCO dataset, which contains segmentation annotations of 80 object classes on 164K images, with additional 91 stuff classes.

Text Embedding. We follow the same process to construct text embeddings as Gu *et al*. [19]. Specifically, we feed prompt engineered texts into the text encoder of CLIP with 85 prompt templates, such as *there is a* {*class name*} *in the scene*, and average the resulting 85 text embeddings of the same class.

Implementation Details. We implement our method on the *MMSegmentation*[4] codebase and inherit its training configurations. Input resolutions are set as 512x512. When using ViT, the pre-trained positional embeddings adopt bicubic interpolation. MaskCLIP requires no training and we train MaskCLIP+ on 4 Tesla V100 GPUs with a batch size of 16. For annotation-free segmentation, we perform MaskCLIP-guided learning for 4k/8k iterations on PASCAL Context/COCO Stuff with DeepLabv2-ResNet101 as the backbone segmentor. Self-training is not used in this setting. For zero-shot segmentation, we choose the lightest training schedule provided by *MMSegmentation*, which is 20k/40k/80k for PASCAL VOC/PASCAL Context/COCO Stuff. The first 1/10 training iterations adopt MaskCLIP-guided learning and the rest adopts self-training. For fair comparisons, we choose DeepLabv2 as the target model for PASCAL VOC

[4] `https://github.com/open-mmlab/mmsegmentation`

Table 1: **Annotation-free segmentation (mIoU). (a)** We evaluate the performance of MaskCLIP(+) on two standard datasets. For Pascal Context, we ignore the evaluation on the background class. The target model of MaskCLIP+ is Deeplabv2-ResNet101. KS and PD denote key smoothing and prompt denoising respectively. And they are not used in MaskCLIP+. **(b)** We test the robustness of MaskCLIP on Pascal Context under various types of corruption

(a)

Method	CLIP	PASCAL Context	COCO Stuff
Baseline	r50	8.3	4.6
	vit16	9.0	4.3
MaskCLIP	r50	18.5	10.2
	+KS	21.0	12.4
	+PD	19.0	10.8
	+KS+PD	21.8	12.8
	vit16	21.7	12.5
	+KS	23.9	13.8
	+PD	23.1	13.2
	+KS+PD	25.5	14.6
MaskCLIP+	r50	23.9	13.6
	vit16	31.1	18.0

(b)

Corruption	**level 1**		**level 5**	
	r50	vit16	r50	vit16
None	18.5	21.7	18.5	21.7
Gaussian Noise	13.7	19.6	2.1	6.8
Shot Noise	14.0	19.6	2.4	7.5
Impulse Noise	9.9	17.3	2.1	7.2
Speckle Noise	15.1	20.0	5.6	11.4
Gaussian Blur	17.4	21.6	4.3	14.1
Defocus Blur	15.7	20.8	6.6	15.5
Spatter	17.1	20.5	7.8	12.2
JPEG	15.7	20.8	7.6	14.5

and COCO Stuff and DeepLabv3+ for PASCAL Context. All use the ResNet-101 backbone initialized with the ImageNet pre-trained weights. Finally, we use the publicly available CLIP-ResNet-50 and CLIP-ViT-B/16 models[5].

4.1 Annotation-Free Segmentation

In this challenging setting, no annotation is provided during training. We first evaluate the mIoU performance on two standard datasets, PASCAL Context and COCO-Stuff. Then we collect images from Flickr to show interesting qualitative results on novel concepts, such as *Batman* and *Joker*. Finally, we test the robustness of MaskCLIP under various image corruptions.

Performance on Standard Datasets. In Table 1a, we show mean Intersection over Union (mIoU) results on PASCAL Context and COCO-Stuff. The baseline in the table refers to directly using dense features from the CLIP's image encoder without any modification. As shown in the table, MaskCLIP outperforms the baseline by huge margins, indicating it is essential to avoid computing attention of the last attention layer and instead value features should

[5] `https://github.com/openai/CLIP`

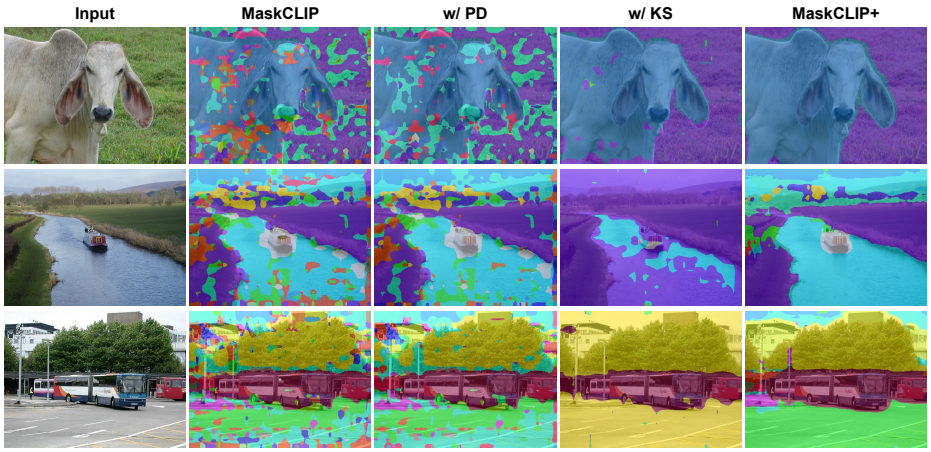

Fig. 3: **Qualitative results on PASCAL Context.** Here all results are obtained **without** any annotation. PD and KS refer to prompt denoising and key smoothing respectively. With PD, we can see some distraction classes are removed. KS is more aggressive. Its outputs are much less noisy but are dominated by a small number of classes. Finally, MaskCLIP+ yields the best results

be directly used. The results also show that key smoothing and prompt denoising are effective and are orthogonal to each other. Therefore, we empirically conclude that for each spatial location, the query features should be discarded and key/value features can be re-organized into final predictions. Furthermore, with the predictions of MaskCLIP as pseudo labels, MaskCLIP+ significantly improves the performance, *e.g.*, on PASCAL Context, without any human annotation, MaskCLIP+(ViT-B/16) obtains 31.1 mIoU. One may notice that ViT almost consistently surpasses ResNet. Apart from ViT-B/16 has more FLOPs than ResNet-50, another possible reason is that ViT only downsamples the input by 16 times whereas ResNet downsamples 32 times, which particularly matters for dense prediction tasks. Besides quantitative results, in Figure 3, we also visualize the outputs of each MaskCLIP variant.

Open-Vocabulary Segmentation on Web-Crawled Images. MaskCLIP inherits the open-vocabulary ability from CLIP and does not require annotations. Therefore, we can deploy it on several interesting setups where the target classes are (1) more fine-grained, such as *red car, yellow car*; (2) of certain imagery properties, *e.g.*, blurry; (3) novel concepts like *Batman, Joker*. To this end, we collect images from Flickr then directly evaluate these images on MaskCLIP and train MaskCLIP+ with only MaskCLIP-guided learning. Note that, for the background, we enumerate a set of classes that might appear in the background and regard them as a whole as the background class. Results in Figure 4 are impressive given the open-vocabulary targets and being annotation-free. Besides, results from MaskCLIP+ are less noisy and more accurate than MaskCLIP, which is complementary to the quantitative results.

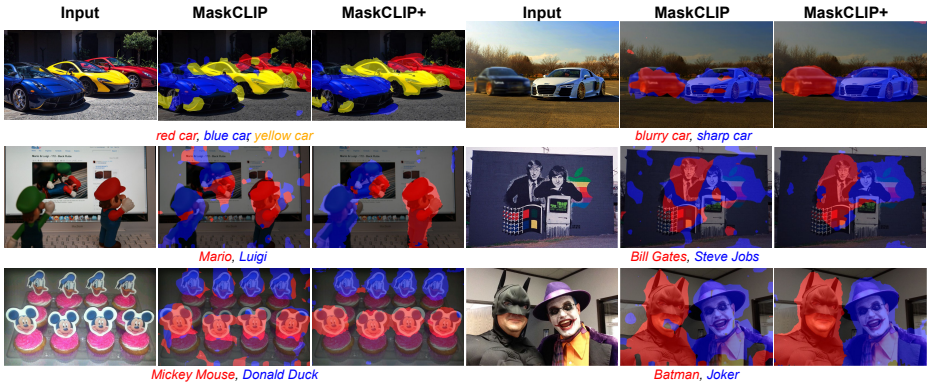

Fig. 4: **Qualitative results on Web images.** Here we show the segmentation results of MaskCLIP and MaskCLIP+ on various **unseen classes**, including fine-grained classes such as cars in different colors/imagery properties, celebrities, and animation characters. All results are obtained **without** any annotation

Robustness Under Corruption. CLIP is trained on web-curated images, whose quality and distribution are more diverse than well-pre-processed datasets. Radford *et al*. [45] and Ravula *et al*. [47] demonstrate the robustness of CLIP on natural distribution shift and artificial corruption respectively. While these explorations are done for image classification, we benchmark its robustness for dense prediction tasks. Specifically, we impose various corruptions used in ImageNet-C [24] with different severity levels on images in PASCAL Context and evaluate on MaskCLIP. In Table 1b, MaskCLIP models based on CLIP-ViT-B/16 are much more robust than CLIP-ResNet-50. In particular, CLIP-ViT-B/16 rarely suffers from degradation across a wide range of corruptions with level 1 severity and is cable of generating reasonable labels even under the most severe corruptions (level 5[6]).

4.2 Zero-Shot Segmentation

Apart from annotation-free segmentation, MaskCLIP+ can also be applied to the zero-shot segmentation task with minor effort. Specifically, in the zero-shot setting, pixels of certain classes do not have annotations, to which MaskCLIP can assign reliable pseudo labels.

Zero-shot Setups. Traditionally, zero-shot segmentation methods train on a subset of classes, named seen classes, with ground-truth annotations, and during inference, both seen and unseen classes are evaluated. Depending on whether the unlabeled pixels are observed during training, the setting can be split into inductive (not observed) and transductive (observed). Our method conforms to the transductive setting.

[6] The severity level is controlled by certain coefficients, such as kernel size, specified in ImageNet-C [24].

Table 2: **Zero-shot segmentation performances.** ST stands for self-training. mIoU(U) denotes mIoU of the unseen classes. SPNet-C is the SPNet with calibration. On PASCAL Context, all methods use DeepLabv3+-ResNet101 as the backbone segmentation model and the rest two datasets use DeepLabv2-ResNet101. For MaskCLIP+, CLIP-ResNet-50 is used to generate pseudo labels

Method	**PASCAL-VOC**			**COCO-Stuff**			**PASCAL-Context**		
	mIoU(U)	mIoU	hIoU	mIoU(U)	mIoU	hIoU	mIoU(U)	mIoU	hIoU
Inductive									
SPNet	0.0	56.9	0.0	0.7	31.6	1.4	·	·	·
SPNet-C	15.6	63.2	26.1	8.7	32.8	14.0	·	·	·
ZS3Net	17.7	61.6	28.7	9.5	33.3	15.0	12.7	19.4	15.8
CaGNet	26.6	65.5	39.7	12.2	33.5	18.2	18.5	23.2	21.2
Transductive									
SPNet+ST	25.8	64.8	38.8	26.9	34.0	30.3	·	·	·
ZS3Net+ST	21.2	63.0	33.3	10.6	33.7	16.2	20.7	26.0	23.4
CaGNet+ST	30.3	65.8	43.7	13.4	33.7	19.5	·	·	·
STRICT	35.6	70.9	49.8	30.3	34.9	32.6	·	·	·
MaskCLIP+	**86.1**	**88.1**	**87.4**	**54.7**	**39.6**	**45.0**	**66.7**	**48.1**	**53.3**
	+50.5	+17.2	+37.6	+24.4	+4.7	+12.4	+46.0	+22.1	+29.9
Fully Sup.	·	88.2	·	·	39.9	·	·	48.2	·

The selection of seen classes varies among previous works and we follow the most common setups, where for PASCAL VOC, the background class is ignored and *potted plant, sheep, sofa, train, tv monitor* are chosen as the 5 unseen classes; for PASCAL Context, the background is not ignored and *cow, motorbike, sofa, cat, boat, fence, bird, tv monitor, keyboard, aeroplane* are unseen; and for COCO Stuff, *frisbee, skateboard, cardboard, carrot, scissors, suitcase, giraffe, cow, road, wall concrete, tree, grass, river, clouds, playing field* are unseen. We report the mean Intersection over Union (mIoU) of seen, unseen, and all classes as well as the harmonic mean (hIoU) of seen and unseen mIoUs as evaluation metrics.

We compare MaskCLIP+ with SOTA methods including SPNet [50], ZS3Net [1], CaGNet [20], and STRICT [41]. ZS3Net and CaGNet are generative approaches, while SPNet is non-generative and more simple but requires post-possessing step of calibration (SPNet-C). STRICT improves SPNet by a self-training strategy and is free of calibration. Compare with these methods, our MaskCLIP+ does not rely on any particular network architecture nor post-possessing. Note that similar to CLIP, all methods, except for ZS3Net, do not exclude unseen classes during pre-training. Besides, MaskCLIP+ also follows the rule that pixel-level annotations of unseen classes are prohibited. Thus, the comparison is fair.

Despite being simple, MaskCLIP+ achieves a strikingly good result. As shown in Table 2, it surpasses all methods on all datasets with large margins. On

Table 3: **Ablations of MaskCLIP+.** Experiments are performed on the PASCAL VOC dataset under the zero-shot setting

Method	mIoU(S)	mIoU(U)	mIoU	hIoU
Adapted DeepLabv2	83.4	3.7	63.5	7.0
+ MaskCLIP-Guided	**89.5**	72.8	85.3	80.3
+ Self-Training	88.8	**86.1**	**88.1**	**87.4**

PASCAL VOC, PASCAL Context, and COCO Stuff, in terms of unseen mIoUs, MaskCLIP+ improves the previous SOTA by 50.5, 24.4, and 46.0 respectively (on a scale of 100). Note that the overall mIoU of MaskCLIP+ is on par with that of fully supervised baselines. Please refer to Table 2 for more specific numbers.

Ablation Studies of MaskCLIP+. We perform ablation studies on the PASCAL VOC zero-shot segmentation setting. As shown in Table 3, we first examine the two proposed strategies in MaskCLIP+. Compared to the adapted DeepLabv2, whose classifier is replaced with the MaskCLIP classifier, MaskCLIP-guided learning improves the unseen mIoU from 3.7 to 72.8 and the result is further improved by self-training to 86.1. However, there is a slight degradation on seen classes when using self-training (from 89.5 to 88.8) partially due to model drifting. Overall, MaskCLIP+ performs better than MaskCLIP on unseen classes and surpasses the baseline DeepLabv2 on seen classes in the same time.

5 Conclusion

This paper presents our exploration of applying CLIP in semantic segmentation, as an early attempt that studies the applicability of pre-trained visual-language models in pixel-level dense prediction tasks. While the conventional fine-tuning paradigm fails to benefit from CLIP, we find the image encoder of CLIP already possesses the ability to directly work as a segmentation model. The resulting model, termed MaskCLIP, can be readily deployed on various semantic segmentation settings without re-training. On top of the success of MaskCLIP, we further propose MaskCLIP+ that leverages MaskCLIP to provide training-time pseudo labels for unlabeled pixels, which thus can be applied to more segmentation-tailored architectures beyond just the image encoder of CLIP. On standard transductive zero-shot segmentation benchmarks, MaskCLIP+ significantly improves previous SOTA results. More importantly, MaskCLIP+ can be readily employed for segmenting more challenging unseen classes, such as celebrities and animation characters.

Acknowledgement. This study is supported under 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). This study is also supported by Singapore MOE AcRF Tier 2 (MOE-T2EP20120-0001) and Shanghai AI Laboratory.

References

1. Bucher, M., Vu, T.H., Cord, M., Pérez, P.: Zero-shot semantic segmentation. In: NeurIPS (2019)
2. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: CVPR (2018)
3. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: NeurIPS (2020)
4. Chen, L.C., Lopes, R.G., Cheng, B., Collins, M.D., Cubuk, E.D., Zoph, B., Adam, H., Shlens, J.: Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In: ECCV (2020)
5. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
6. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
8. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: CVPR (2021)
9. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint (2015)
10. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR (2021)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
12. Desai, K., Johnson, J.: Virtex: Learning visual representations from textual annotations. In: CVPR (2021)
13. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV (2015)
14. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. In: NeurIPS (2014)
15. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. IJCV (2015)
16. Gomez, L., Patel, Y., Rusiñol, M., Karatzas, D., Jawahar, C.: Self-supervised learning of visual features through embedding images into text topic spaces. In: CVPR (2017)
17. Gordo, A., Larlus, D.: Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In: CVPR (2017)
18. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., et al.: Bootstrap your own latent: A new approach to self-supervised learning. In: NeurIPS (2020)
19. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint (2021)
20. Gu, Z., Zhou, S., Niu, L., Zhao, Z., Zhang, L.: Context-aware feature generation for zero-shot semantic segmentation. In: ACM MM (2020)
21. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: ICCV (2011)
22. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)

23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
24. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: ICLR (2019)
25. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: EMNLP (2021)
26. Hu, P., Sclaroff, S., Saenko, K.: Uncertainty-aware learning for zero-shot semantic segmentation. In: NeurIPS (2020)
27. Hung, W.C., Tsai, Y.H., Liou, Y.T., Lin, Y.Y., Yang, M.H.: Adversarial learning for semi-supervised semantic segmentation. In: BMVC (2018)
28. Ibrahim, M.S., Vahdat, A., Ranjbar, M., Macready, W.G.: Semi-supervised semantic image segmentation with self-correcting networks. In: CVPR (2020)
29. Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Label propagation for deep semi-supervised learning. In: CVPR (2019)
30. Jain, A., Tancik, M., Abbeel, P.: Putting nerf on a diet: Semantically consistent few-shot view synthesis. In: ICCV (2021)
31. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021)
32. Kato, N., Yamasaki, T., Aizawa, K.: Zero-shot semantic segmentation via variational mapping. In: ICCVW (2019)
33. Lee, D.H.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: ICMLW (2013)
34. Li, P., Wei, Y., Yang, Y.: Consistent structural relation learning for zero-shot segmentation. In: NeurIPS (2020)
35. Li, X., Sun, Q., Liu, Y., Zhou, Q., Zheng, S., Chua, T.S., Schiele, B.: Learning to self-train for semi-supervised few-shot classification. In: NeurIPS (2019)
36. Mendel, R., De Souza, L.A., Rauber, D., Papa, J.P., Palm, C.: Semi-supervised segmentation based on error-correcting supervision. In: ECCV (2020)
37. Mittal, S., Tatarchenko, M., Brox, T.: Semi-supervised semantic segmentation with high-and low-level consistency. IEEE TPAMI (2019)
38. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: CVPR (2014)
39. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: ECCV (2016)
40. Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: CVPR (2020)
41. Pastore, G., Cermelli, F., Xian, Y., Mancini, M., Akata, Z., Caputo, B.: A closer look at self-training for zero-label semantic segmentation. In: CVPRW (2021)
42. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: ICCV (2021)
43. Pham, H., Dai, Z., Xie, Q., Le, Q.V.: Meta pseudo labels. In: CVPR (2021)
44. Quattoni, A., Collins, M., Darrell, T.: Learning visual representations using images with captions. In: CVPR (2007)
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
46. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. arXiv preprint (2021)

47. Ravula, S., Smyrnis, G., Jordan, M., Dimakis, A.G.: Inverse problems leveraging pre-trained contrastive representations. In: NeurIPS (2021)
48. Sariyildiz, M.B., Perez, J., Larlus, D.: Learning visual representations with caption annotations. In: ECCV (2020)
49. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
50. Xian, Y., Choudhury, S., He, Y., Schiele, B., Akata, Z.: Semantic projection network for zero-and few-label semantic segmentation. In: CVPR (2019)
51. Xie, J., Zheng, S.: Zsd-yolo: Zero-shot yolo detection using vision-language knowledge distillation. arXiv preprint (2021)
52. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: NeurIPS (2014)
53. Yuan, X., Lin, Z., Kuen, J., Zhang, J., Wang, Y., Maire, M., Kale, A., Faieta, B.: Multimodal contrastive training for visual representation learning. In: CVPR (2021)
54. Zhao, H., Puig, X., Zhou, B., Fidler, S., Torralba, A.: Open vocabulary scene parsing. In: ICCV (2017)
55. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)
56. Zou, Y., Zhang, Z., Zhang, H., Li, C.L., Bian, X., Huang, J.B., Pfister, T.: Pseudoseg: Designing pseudo labels for semantic segmentation. In: ICLR (2021)