Extract Free Dense Labels from CLIP Supplementary Materials

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

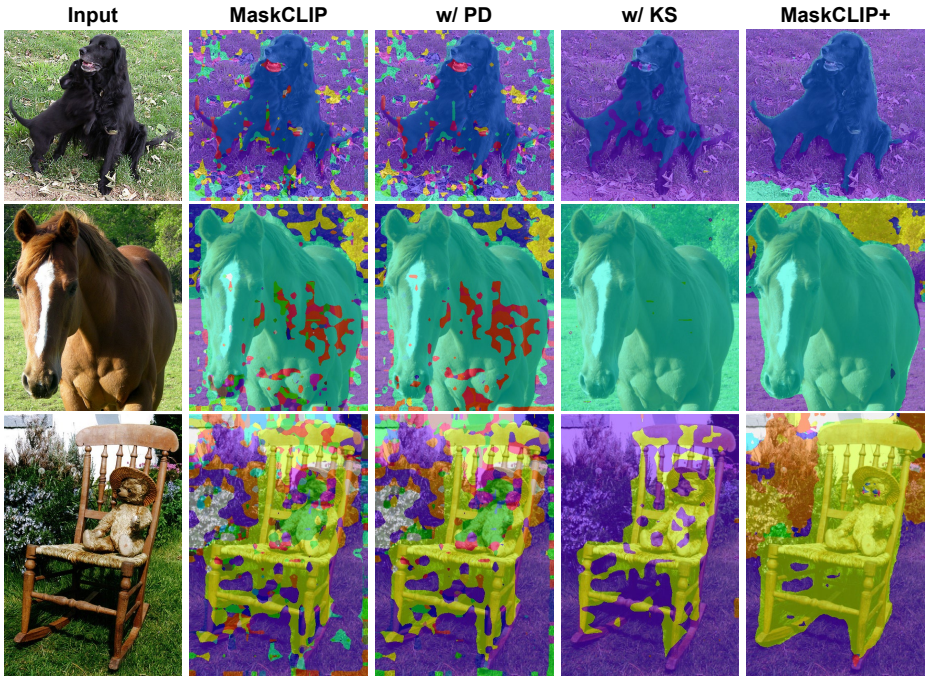

Fig. 1: **More qualitative results on PASCAL Context.** Here all results are obtained **without** any annotation. PD and KS refer to prompt denoising and key smoothing respectively. Row 2, Col 4 shows a failure case of KS, where all the pixels in the image are labeled as the *horse*. Note that, PASCAL Context does not contain *bear* or *teddy bear* classes and MaskCLIP predicts the *teddy bear* pixels as *bedclothes*

1 Qualitative Results on Annotation-Free Segmentation

In Figure 1, we show more qualitative results of MaskCLIP on the PASCAL Context dataset in the annotation-free setting. The results are consistent with

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

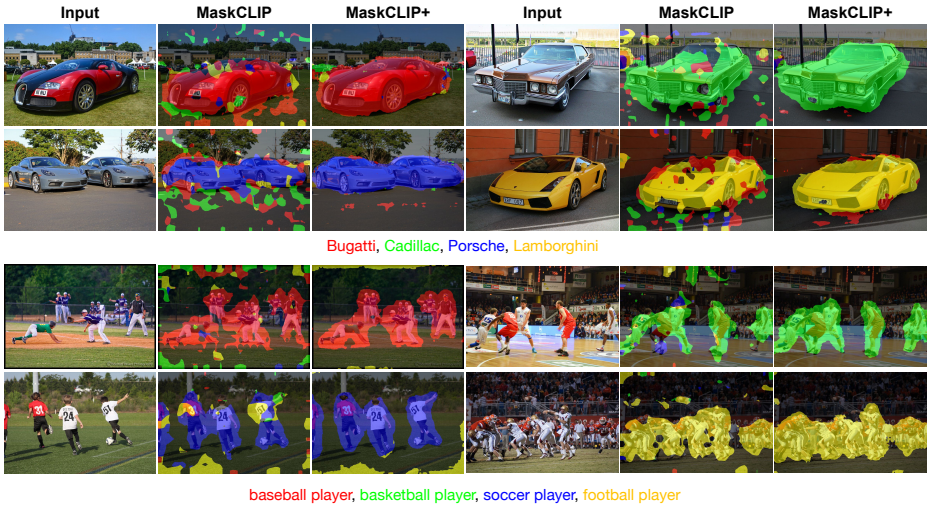

Fig. 2: **More qualitative results on Web images.** MaskCLIP and MaskCLIP+ can yield reasonable segmentation results of different car brands and sports **without** any annotation

our analysis in the main submission, where prompt denoising (PD) removes the unconfident distraction classes, key smoothing (KS) aggressively smooths the noisy predictions, and MaskCLIP+ yields the best results through pseudo-label-training. We find the predictions of KS are often dominated by a few classes and we show a failure case in the Figure 1 (Row 2, Col 4), where one class dominates the whole image. Moreover, the behavior of MaskCLIP and MaskCLIP+ in Row 3 is interesting. Since the PASCAL Context dataset does not contain *bear* or *teddy bear* classes, MaskCLIP classifies the *teddy bear* pixels into *bedclothes*, which is the most related class. Meanwhile, through pseudo-label-training, after observing the true *bedclothes* pixels, MaskCLIP+ decides to treat the teddy bear as part of the chair that it sits on.

In our main submission, we show qualitative results of fine-grained classes (*red car, yellow car*), objects with certain imagery properties (*blurry car*), and novel concepts (*Batman, Bill Gates*). Since MaskCLIP preserves the open-vocabulary ability, we can evaluate it on many interesting setups. In Figure 2 we test whether MaskCLIP can segment out different car brands and sports. Similar to our main submission, the evaluation images are crawled from Flickr and all results are obtained without any annotation. MaskCLIP and MaskCLIP+ again demonstrate powerful open-vocabulary ability on subtle concepts. Note that, in the *basketball* and *football* examples, MaskCLIP not only correctly distinguishes athletes playing different sports, but also separates audience and players.

2 Robustness Results on Annotation-Free Segmentation

In our main submission, we test the robustness of MaskCLIP under artificial corruptions. We use corrupting operations provided by the official code of ImageNet-

Table 1: **More robustness results.** Here we evaluate MaskCLIP on PASCAL Context in the annotation-free setting under ImageNet-C corruptions across all severity levels. Results are reported in the mIoU metric

Corruption	**level 1**		**level 2**		**level 3**		**level 4**		**level 5**	
	r50	vit16	r50	vit16	r50	vit16	r50	vit16	r50	vit16
None	18.5	21.7	18.5	21.7	18.5	21.7	18.5	21.7	18.5	21.7
Gaussian Noise	13.7	19.6	11.2	17.7	7.9	14.8	4.7	11.1	2.1	6.8
Shot Noise	14.0	19.6	11.0	17.6	7.8	14.8	4.0	10.4	2.4	7.5
Impulse Noise	9.9	17.3	8.1	15.9	6.7	14.4	4.1	10.9	2.1	7.2
Speckle Noise	15.1	20.0	13.6	19.0	9.6	16.0	7.6	14.0	5.6	11.4
Gaussian Blur	17.4	21.6	14.4	20.4	11.1	18.9	8.1	17.3	4.3	14.1
Defocus Blur	15.7	20.8	14.0	20.1	10.9	18.6	8.5	17.1	6.6	15.5
Spatter	17.1	20.5	13.0	17.9	10.9	16.4	10.1	14.5	7.8	12.2
JPEG	15.7	20.8	14.3	20.1	13.3	19.5	10.3	17.4	7.6	14.5

Table 2: **Baselines for Robustness Test.** Evaluation on PASCAL Context under level 5 corruptions with the ViT-B/16 backbone. N.: Noise, B.: Blur

	None	Gauss N.	Shot	Impulse	Speckle	Gauss B.	Defocus	Spatter	JPEG
MaskCLIP	21.7	**6.8**	**7.5**	**7.2**	11.4	14.1	15.5	12.2	14.5
Fully Sup.	**54.5**	5.1	6.7	4.8	**22.7**	**37.1**	**40.1**	**31.5**	**39.8**

C [1]. In particular, the severity levels are controlled by a series of coefficients of corruption operators. Limited by space, in the main submission, we only include results of level 1 and level 5. Here, we extend the table to all levels. As shown in Table 1, CLIP-ViT-B/16 consistently outperforms CLIP-ResNet-50 by large margins and shows decent robustness.

We also supplement a baseline for the robustness test to compare with Table 1(b) in our main submission. In particular, we train an FCN segmentation model with the ViT-B/16 backbone (initialized with ImageNet-21K pre-trained weights) on PASCAL Context in a fully supervised manner for 40K iterations, then test the model on corrupted inputs. Table 2 shows that MaskCLIP performs particularly well on Gaussian/shot/impulse noises.

3 Quantitative Results on Zero-Shot Segmentation

In Table 3, we report the mIoUs on seen classes of various methods. As mentioned in our main submission, across three standard datasets, using pseudo labels as the guidance, instead of distillation by feature matching, does not affect MaskCLIP+'s performance on seen classes.

Table 3: **Zero-shot segmentation performances on seen classes (mIoU)**

Method	**PASCAL-VOC**	**COCO-Stuff**	**PASCAL-Context**
SPNet	75.8	34.6	·
SPNet-C	78.0	35.2	·
ZS3Net	77.3	34.7	20.8
CaGNet	78.4	35.5	24.8
SPNet	77.8	34.6	·
ZS3Net	78.0	34.9	27.0
CaGNet	78.6	35.6	·
STRICT	82.7	35.3	·
MaskCLIP+	**88.8** (+6.1)	**38.2** (+2.9)	**44.4** (+17.4)
Fully Sup.	88.6	38.1	44.4

Table 4: **Zero-shot segmentation performances (pAcc & mAcc)**

Method	**PASCAL-VOC**			**COCO-Stuff**			**PASCAL-Context**		
	$pAcc_{(S)}$	$pAcc_{(U)}$	pAcc	$pAcc_{(S)}$	$pAcc_{(U)}$	pAcc	$pAcc_{(S)}$	$pAcc_{(U)}$	pAcc
SPNet	**94.8**	0.0	76.9	65.6	1.7	51.3	·	·	·
SPNet-C	88.8	29.6	77.6	61.8	24.5	53.4	·	·	·
ZS3Net	93.0	21.5	79.4	64.3	22.8	54.7	53.5	58.6	52.8
CaGNet	89.5	43.0	80.7	65.6	25.5	56.6	55.2	66.8	56.6
ZS3Net	91.9	34.1	81.0	65.8	24.9	56.3	46.8	70.2	49.5
CaGNet	87.0	58.6	81.6	**65.9**	26.7	56.8	·	·	·
MaskCLIP+	94.6 (-0.2)	**91.4** (+32.8)	**94.0** (+12.4)	64.2 (-1.7)	**79.4** (+52.7)	**67.6** (+10.8)	**73.9** (+18.7)	**82.3** (+12.1)	**74.8** (+18.2)
Fully Sup.	·	·	94.0	·	·	68.1	·	·	74.8

Method	**PASCAL-VOC**			**COCO-Stuff**			**PASCAL-Context**		
	$mAcc_{(S)}$	$mAcc_{(U)}$	mAcc	$mAcc_{(S)}$	$mAcc_{(U)}$	mAcc	$mAcc_{(S)}$	$mAcc_{(U)}$	mAcc
SPNet	**94.6**	0.0	70.9	50.3	0.0	45.9	·	·	·
SPNet-C	87.9	23.9	71.9	46.3	16.1	43.6	·	·	·
ZS3Net	87.7	15.8	73.5	50.4	27.0	48.4	23.8	43.2	27.0
CaGNet	88.7	39.4	76.4	50.7	27.0	48.5	35.7	49.8	36.8
ZS3Net	85.7	26.4	73.8	50.4	27.2	48.6	32.3	57.1	36.4
CaGNet	83.9	50.7	75.6	50.6	27.3	48.5	·	·	·
MaskCLIP+	93.7 (-0.9)	**92.6** (+41.9)	**93.4** (+17.0)	**50.8** (+0.1)	**72.4** (+45.1)	**52.7** (+4.1)	**55.4** (+19.7)	**80.0** (+22.9)	**59.5** (+22.7)
Fully Sup.	·	·	93.4	·	·	53.0	·	·	59.5

Fig. 3: **Open-vocabulary segmentation with a larger target text set**

Apart from Intersection over Union (IoU), some zero-shot segmentation methods also report pixel accuracy (pAcc) and mean accuracy (mAcc) as evaluation metrics. For comprehensive comparisons, we provide performance with the mentioned metrics in Table 4. In terms of the overall and unseen pAcc/mAcc, MaskCLIP+ still surpasses the previous SOTA methods by large margins and reaches near the fully-supervised baselines. However, its pAcc/mAcc of seen classes on PASCAL VOC and COCO-Stuff fall behind SPNet and CaGNet+ST by a bit. Different from mIoU, pAcc and mAcc punish only false negatives but not false positives (mIoU punishes both). Previous methods are much more confident on seen classes than unseen classes, therefore yields more predictions on seen classes, which consequently avoids false negatives on seen classes. In fact, SPNet biases towards seen classes so much that without calibration (reduce the confidence of seen classes by scaling factors), its performance on unseen classes is almost zero. MaskCLIP+, on the contrary, is more balanced between seen and unseen classes.

4 Vocabulary Used in Open-Vocabulary Segmentation

In Figure 1 and Figure 4 of our main submission, all images share the same background classes, i.e., *building, ground, grass, tree, sky*. For foreground classes, different images have a different set of targets, which are shown right below each image in Figure 4. In Figure 3, we supplement an example with a larger vocabulary, with *Batman, Joker, James Gordon, The Penguin, Robin, Alfred, Catwoman, Harley Quinn* as the foreground and all classes except *person* in the Cityscapes as the background. We observe that Batman's jaw is segmented as *James Gordon* and part of Joker's suit is classified into *The Penguin*. Since certain local features are shared among multiple characters, it reveals that sometimes MaskCLIP cannot see broadly enough.

5 Input Resolution and Multi-Scale Ensemble

There is a trade-off in terms of the input resolution of MaskCLIP. Using the same input resolution as CLIP (224x224) assures the resolution/positional encoding matching but at the cost of yielding smaller output. We empirically find there

Table 5: **Input resolutions and multi-scale ensemble.** Here, we evaluate MaskCLIP on the PASCAL Context dataset

Input Res.	224	336	520	[224, 520]	[224, 336, 520]
mIoU	22.72	23.02	21.68	25.16	**26.34**

Algorithm 1: MaskCLIP+ pseudo code

$P \leftarrow$ MaskCLIP model;
$T \leftarrow$ text embeddings of target classes;
$V_1 \leftarrow$ target model initialized w/ IN pre-trained;
$V_1 \leftarrow$ load T to classifier weights of V_1;
$\mathcal{D} \leftarrow$ images for training;
$N_g \leftarrow$ MaskCLIP-guided learning iterations;
$N_s \leftarrow$ self-training iterations;
for $i = 1, 2, \ldots, N_g$ **do**
  $\hat{y} \leftarrow$ model prediction $V_i(\mathcal{D}_i)$;
  $y \leftarrow$ pseudo labels from MaskCLIP $P(\mathcal{D}_i)$;
  $\mathcal{L} \leftarrow$ cross entropy loss $\mathcal{L}_{CE}(\hat{y}, y)$;
  $V_{i+1} \leftarrow$ SGD model update;
end
for $j = N_g + 1, N_g + 2, \ldots, N_g + N_s$ **do**
  $\hat{y} \leftarrow$ model prediction $V_j(\mathcal{D}_j)$;
  $y \leftarrow$ self-generated pseudo labels $V_j(\mathcal{D}_j)$;
  $\mathcal{L} \leftarrow$ cross entropy loss $\mathcal{L}_{CE}(\hat{y}, y)$;
  $V_{j+1} \leftarrow$ SGD model update;
end

exists a sweet spot at 336x336. We also find that multi-scale ensembles mitigate the resolution problem. (See Table 5.)

6 Pseudo Code of MaskCLIP+

The complete training process of MaskCLIP+ is illustrated in Algorithm 1.

References

1. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: ICLR (2019)