Frozen CLIP: A Strong Backbone for Weakly Supervised Semantic Segmentation

Supplementary Material

In the supplementary material, we will show some details about how to generate the initial CAM, the framework for the fully-supervised case and provide more experimental results to verify our WeCLIP.

1. Initial CAM Generation

We follow [3] to generate the initial CAM. For a given image I with class label set C_I, the image is input to the frozen CLIP image encoder to generate the image feature map as $F \in \mathbb{R}^{d\times(hw)}$, after passing global average pooling, the feature vector $F_v \in \mathbb{R} \times 1$ is generated. Meanwhile, The class labels set C_I, with the pre-defined background label set C_{bg} [3], are used to build text prompts using the text "a clear origami $\{*\}$", where $*$ is the specific class label. Then the text prompts are input to the text encoder to generate the feature map $F_t \in \mathbb{R}^{d\times(|C_I|+|C_{\text{bg}}|)}$. Using F_v and F_t, the distance is compute as:

$$D = \frac{F_t F_v^T}{||F_t|| \cdot ||F_v||}, \tag{1}$$

where $D \in \mathbb{R}^{(|C_I|+|C_{\text{bg}}|)\times 1}$.

Then, the distance is passed to the softmax function to generate the class scores:

$$S^c = softmax(D/\tau), \tag{2}$$

where S^c is the classification score for class c, and $c \in \{C_{\text{bg}}, C_I\}$, τ is the temperature parameter.

Using GradCAM [4], we can generate the feature weight map for a specific class c in the kth channel:

$$w_c^k = \frac{1}{hw}\sum_{i=1}^{h}\sum_{j=1}^{w}\sum_{c'}\frac{\partial S^c}{\partial D^{c'}}\frac{\partial D^{c'}}{\partial F_{i,j}^k}, \tag{3}$$

where $c \in \{C_{\text{bg}}, C_I\}$ and $c' \in \{C_{\text{bg}}, C_I\}$.

Finally, the initial CAM for the specific foreground class c is computed as:

$$M_{\text{init}}^c(i,j) = \text{ReLU}(\sum_k w_c^k F_{i,j}^k). \tag{4}$$

For more details, please refer to [3].

2. More Experimental Results

To show the effectiveness of our approach, we compare the quality of the pseudo labels with other multi-stage approaches in Tab. 1. Since our WeCLIP is a single-stage solution, we directly use segmentation predictions as the pseudo labels for comparison. In other words, by using the prediction as the pseudo labels, our approach can be regarded as a pseudo label generation part of the multi-stage solution, which aims to provide high-quality pseudo labels to train an individual segmentation model. It can be seen that our approach significantly outperforms other approaches. For example, compared to the CLIP-based solutions such as CLIMS [6] and CLIP-ES [3], our approach brings out more than 3% mIoU increase. Fig. 1 shows some qualitative comparisons, which also illustrates our approach can generate high-quality pseudo labels. Ours are more complete and smooth.

Table 1. Performance comparison about the generated pseudo labels between our approach and others on PASCAL VOC 2012 *train* set. Note that we regard WeCLIP as a pseudo label generation method and directly use its predictions as the pseudo labels.

Method	Pub.	Sup.	mIoU(%)
RIB [2]	NeurIPS'21	I	70.6
MCTformer [7]	CVPR'22	I	69.1
ACR [1]	CVPR'23	I	72.3
CLIMS [6]	CVPR'22	I+L	70.5
CLIP-ES [3]	CVPR'23	I+L	75.0
ours-WeCLIP	-	I+L	**78.2**

Table 2. Ablation study of the input frozen image features for decoder on PASCAL VOC 2012 *val* set. "1, 5, 8, 11, 12" indicates the value of N_0. For example, $N_0 = 1$ means that frozen image features from 1 to 12 layers (all layers) are selected as input for the decoder.

$\{F_{\text{init}}^l\}_{l=N_0}^{l=12}$	1	5	8	11	12
mIoU (%)	**74.9**	74.7	74.6	74.5	74.3

In Tab. 2, we conduct the ablation study to illustrate the influence of different frozen image features, which are selected as input for our decoder. When $N_0 = 1$, image features from all blocks in the frozen image encoder are selected, and the best performance is generated. Besides, N_0 from 1 to 12, the mIoU score is decreased from 74.9% to 74.3%, indicating that fewer features are selected, and lower performance is generated. The possible reason is that using all features has a more comprehensive semantic representation.

Tab. 3 is the ablation study for the different supervision signals of A_f. M_p means using the online pseudo labels for

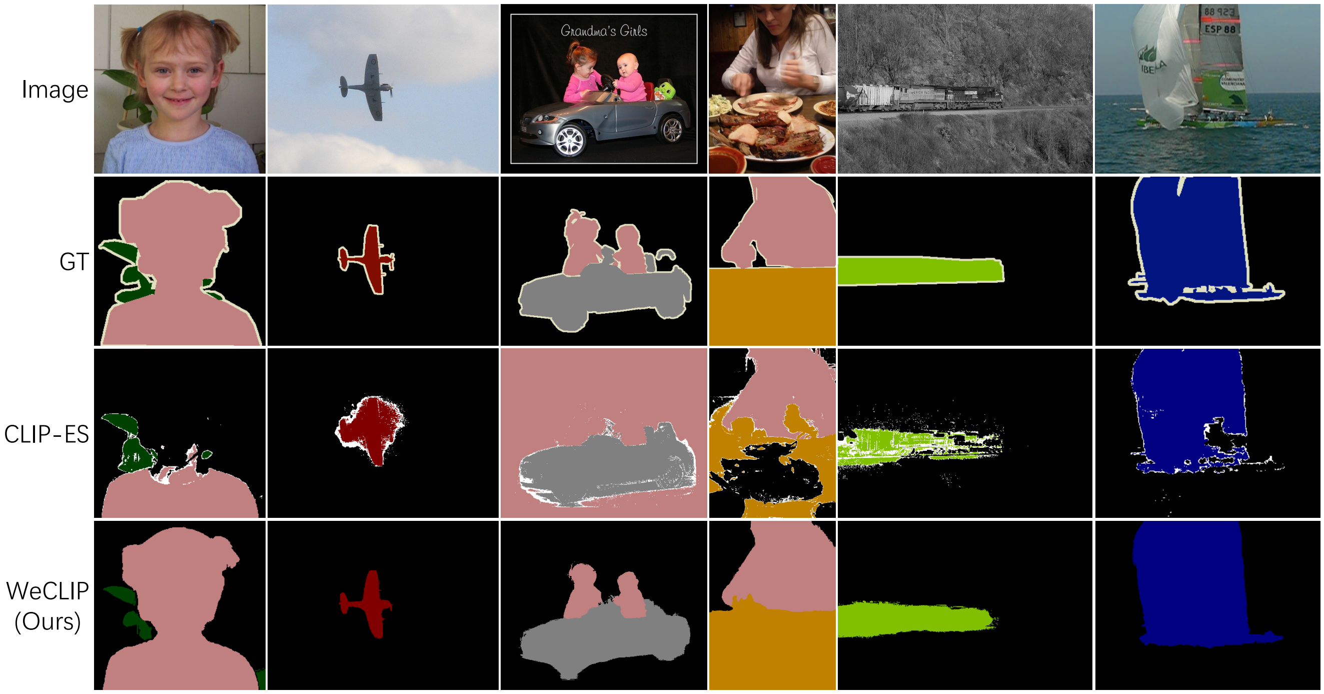

Figure 1. Qualitative comparison about the generated pseudo labels between our approach and CLIP-ES [3] on PASCAL VOC 2012 *train* set. Our approach generates more accurate pseudo labels.

Table 3. Ablation study for the supervision of A_f. M_p is the online pseudo labels, P is the final prediction. simultaneously using M_p and P means using the intersection between M_p and P.

$\hat{A}$		mIoU (%)
M_p	$\text{argmax}(P)$	
✓		**74.9**
	✓	74.6
✓	✓	74.8

$\hat{A}$. $\text{argmax}(P)$ means using the final prediction P for $\hat{A}$. The last row means using the intersection between M_p and P for $\hat{A}$. It can be found that when using the pseudo label M_p to produce $\hat{A}$ as supervision, it achieves 74.9% mIoU, which performs better than the other two cases. Using the prediction P cannot bring a higher mIoU score since P is updated during training, and it is easy to produce conflict supervision, leading to an ineffective learning process.

Table 4. Ablation study of the hyperparameter λ for balancing the loss function.

λ	0	0.1	0.5	1.0
mIoU	73.3	**74.9**	73.6	72.9

Tab. 4 shows the influence of the hyperparameter λ for balancing the loss function. When $\lambda = 0$, the learning of affinity map A_f is not supervised. It only generates a 73.3% mIoU score. This is because the uncontrolled A_f makes the filter G^l and refining map R unstable, thus reducing the quality of online pseudo labels. When $\lambda = 0.1$, it produces better results than others, showing a good balance between two loss functions.

Table 5. Influence of different multi-scales during inference.

Multi-scale	mIoU (%)
{1.0}	74.0
{0.5, 1.0}	74.2
{0.75, 1.0}	**74.9**
{0.5, 0.75, 1.0}	74.4
{0.75, 1.0, 1.25}	74.8
{0.75, 1.0, 1.5}	74.5

Tab. 5 shows the influence of the multi-scale strategy during inference. It can be seen that {0.75, 1.0} performs better than other settings. Introducing a larger scale, such as 1.5, does not improve the performance, showing that the Frozen CLIP backbone is not sensitive to the large scale.

In Fig. 2, we show more feature visualization results to compare the difference between the CLIP features and ImageNet features. For each pair visualization (each column), we randomly select 200 images from the PASCAL VOC 2012 *train* set. All other settings are the same as our paper.

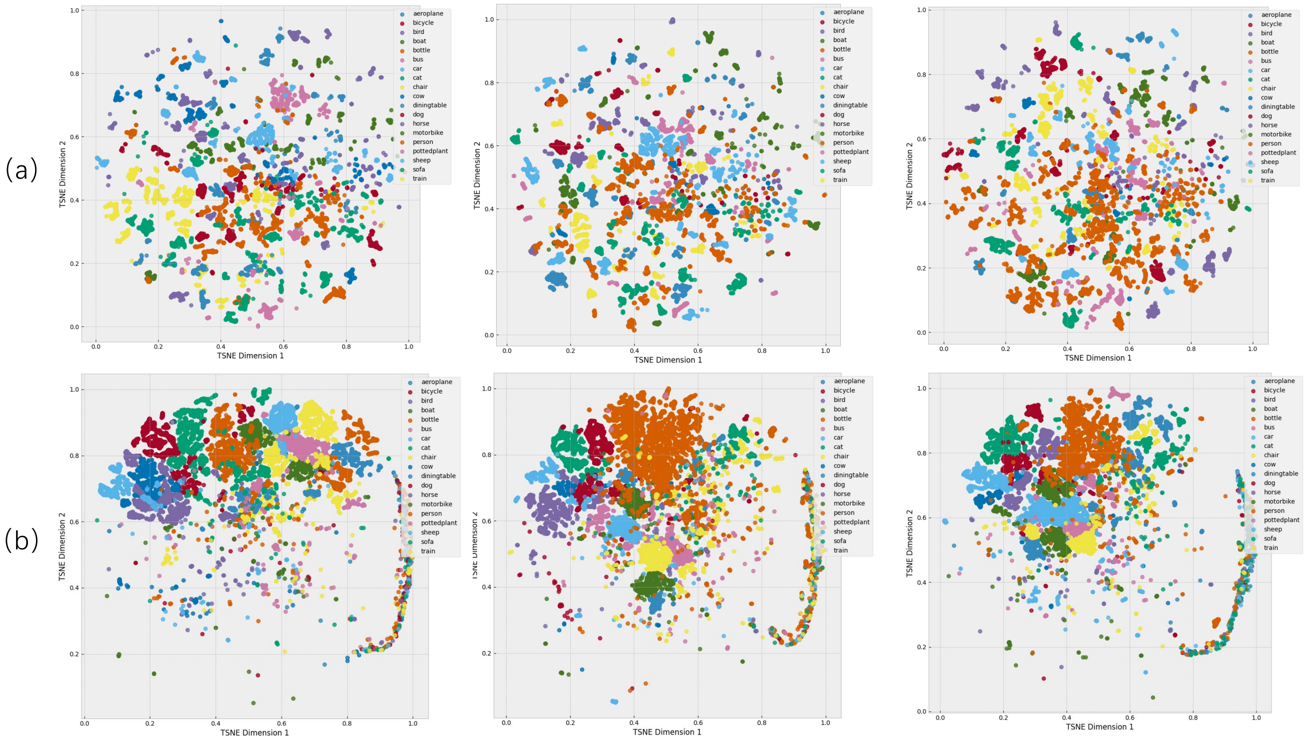

Figure 2. Feature visualization with T-SNE [5] to show why frozen CLIP can be used for semantic segmentation. Each color represents one specific class. (a) Frozen ImageNet pre-trained feature of ViT-B. (b) Frozen CLIP pre-trained vision feature of VIT-B. It can be seen that without any retraining, the features belonging to the same class from the frozen CLIP are denser and more clustered than the ImageNet pre-trained features. Best viewed in color.

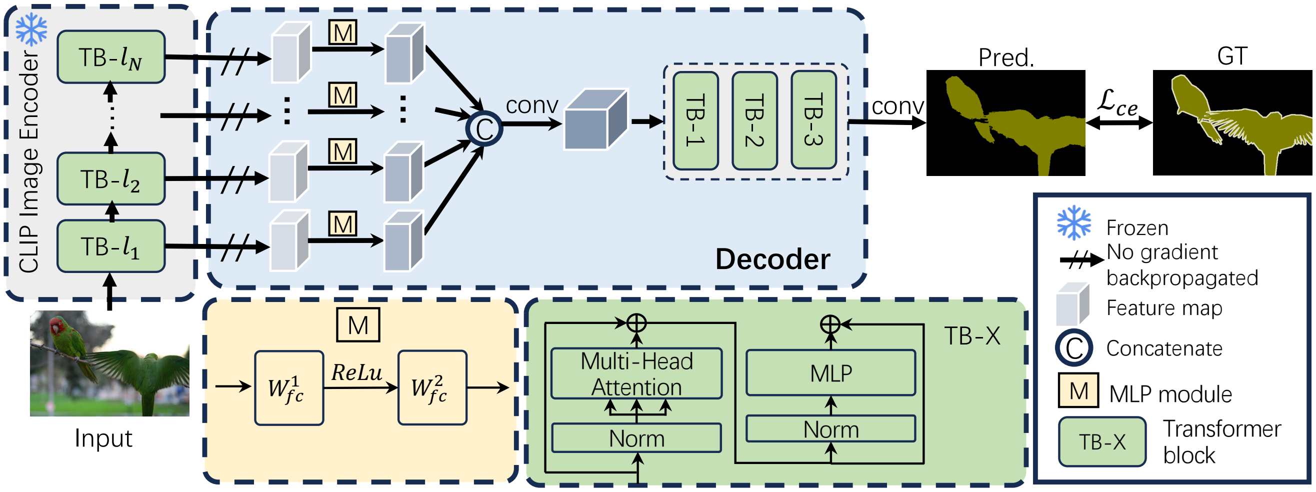

Figure 3. Framework for fully-supervised semantic segmentation. Given an image, it passes the frozen CLIP image encoder to extract the feature map, which is then input to our decoder to generate the final prediction.

It can be found that features belonging to the same class, pre-trained by CLIP, are denser and clustered, while features pre-trained by ImageNet are more sparse and decentralized, which explains why the frozen CLIP feature can be directly used for semantic segmentation. Fig. 2 indicates that the extracted features from the CLIP model can better represent semantic information for different classes, making features belonging to different classes not confused. With such discriminative features, It is more convenient to conduct segmentation tasks.

3. Framework for Fully-supervised Semantic Segmentation

In Fig. 3, we show the framework of our approach for fully-supervised semantic segmentation. We directly use our decoder as the decoder to learn from the provided pixel-level supervision. Our RFM is not used as it is unnecessary to refine the pixel-level supervision.

4. Background Text Set

We follow CLIP-ES [3] to define the background class set. For PASCAL VOC 2012 set, the set is $\{$*ground, land, grass, tree, building, wall, sky, lake, water, river, sea, railway, railroad, keyboard, helmet, cloud, house, mountain, ocean, road, rock, street, valley, bridge, sign*$\}$, For MS COCO-2014, $\{$*sign, keyboard*$\}$ is removed. Besides, the text prompt for the background class is '*a clear origami* $\{$*background class*$\}$'.

References

[1] Hyeokjun Kweon, Sung-Hoon Yoon, and Kuk-Jin Yoon. Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor. In *CVPR*, pages 11329–11339, 2023. 1

[2] Jungbeom Lee, Jooyoung Choi, Jisoo Mok, and Sungroh Yoon. Reducing information bottleneck for weakly supervised semantic segmentation. In *NeurIPS*, pages 27408–27421, 2021. 1

[3] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *CVPR*, pages 15305–15314, 2023. 1, 2, 4

[4] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 1

[5] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008. 3

[6] Jinheng Xie, Xianxu Hou, Kai Ye, and Linlin Shen. Clims: cross language image matching for weakly supervised semantic segmentation. In *CVPR*, pages 4483–4492, 2022. 1

[7] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *CVPR*, pages 4310–4319, 2022. 1