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

- Problem / objective
  - CLIP 사용해서 Open-Vocabulary Semantic Segmentation
- Contribution / Key idea
  - **MaskCLIP**: CLIP 사용해서 training-free + open-vocabulary semantic segmentation한 첫 논문
    - CLIP vision encoder 마지막 layer에서, value embedding을 direct하게 사용 (query, key 제거)
  - **MaskCLIP+**: MaskCLIP을 통해 얻은 pseudo-label로 self-training

ZHOU, Chong; LOY, Chen Change; DAI, Bo. Extract free dense labels from clip. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022. p. 696-712.

ECCV 2022 Oral

- **CLIP의 visual & language features**
- **CLIP features를 pixel-level dense prediction tasks에 사용**
  - 선행 연구: CLIP features를 global image representation로 사용
  - Ours: CLIP features를 object-level and local semantic representation로 사용
- **실험 결과 얻은 결론**
  - (1) 원본 CLIP feature space의 vision-language association을 깨지 말자. (해보니까 CLIP의 image encoder를 segmentation task에 fine-tuning하면 오히려 성능이 떨어짐)
  - (2) CLIP의 text embeddings을 바꾸려는 불필요한 시도하지 않는 것이 중요. (해보니까 unseen classes에 대한 segmentation을 잘 못하게 됨)
- **MaskCLIP**
  - (1) CLIP's image encoder의 dense patch-level features: 최종 attention layer의 value feature
  - (2) CLIP's text encoder의 text embeddings를 direct 사용
  - (3) Dense prediction을 위한 분류 가중치: 1x1 convolutions, directly obtained by text embeddings
  - (4) 2가지 마스크 개선 기법
    - (i) Key smoothing: (마지막 attention layer에서) 서로 다른 패치의 key features 간 유사도를 통해 predictions를 smoothing
    - (ii) Prompt denoising: 이미지에 없을것 같은 클래스 프롬프트 제거함으로서, predictions 정확도 향상 도모

- **Conventional Fine-Tuning Hinders Zero-Shot Ability**
  - ❏ DeepLab에 2가지 CLIP-specific 수정 적용
    1. Backbone: CLIP 이미지 인코더의 사전학습된 가중치로 딥랩 초기화
    2. Mapper: CLIP 텍스트 임베딩을 딥랩 분류기 가중치로 사용 (1x1 convolution layer)
  - ❏ 결과
    Seen classes에 대해서는 성능 good, 그러나 unseen classes에 대해서는 성능 bad
  - ❏ 결과 분석
    Unseen classes에 대해서 성능 안좋아진 이유는, 기존 CLIP features의 visual-language association이 깨져서일 것이다.
    (1) 네트워크 구조적으로, backbone이 CLIP 이미지 인코더와 살짝 다름.
    (2) CLIP image encoder로부터 초기화된 가중치가 파인튜닝하며 계속 업데이트됨.
    (3) Mapper가 seen classes에 대해서만 학습되니까 일반화 성능 악화.
  - ❏ 결론
    파라미터 추가 및 CLIP의 feature space를 수정하려고 하지 말자.

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

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

ZHOU, Chong; LOY, Chen Change; DAI, Bo. Extract free dense labels from clip. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022. p. 696-712.

ECCV 2022 Oral

- **MaskCLIP / MaskCLIP+**

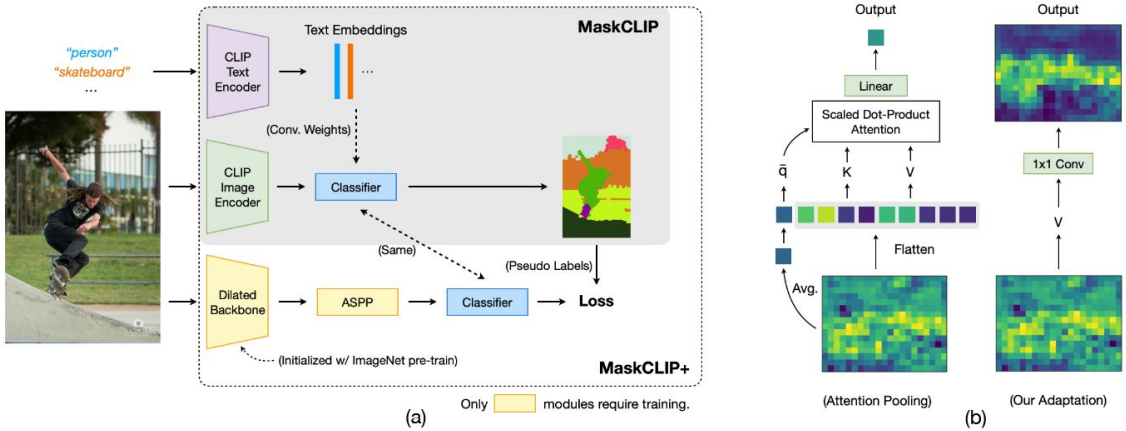


Fig. 2: **Overview of MaskCLIP/MaskCLIP+.** Compared to the conventional fine-tuning method, the key to the success of MaskCLIP is keeping the pre-trained weights frozen and making minimal adaptation to preserve the visual-language association. Besides, to compensate for the weakness of using the CLIP 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]

전유진

ZHOU, Chong; LOY, Chen Change; DAI, Bo. Extract free dense labels from clip. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022. p. 696-712.

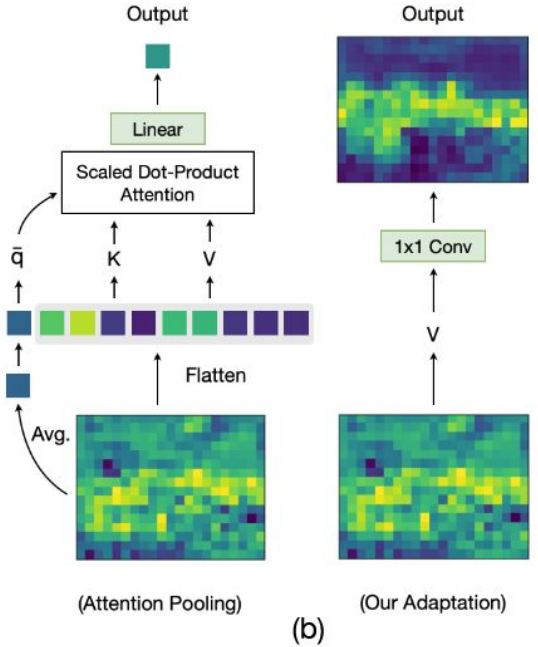
ECCV 2022 Oral

- **MaskCLIP**
  - ❏ CLIP (ResNet-based) 이미지 인코더의 global attention pooling layer
    - 쿼리: globally average-pooled feature / 키, 밸류: spatial local features
    - Transformer layer의 아웃풋을 이미지의 global representation으로 사용.
  - ❏ 가설
    각 위치의 value features가 CLIP text embeddings과 상응하는 풍부한 local 의미 정보를 담고있어서 transformer layer의 아웃풋을 global representation으로 사용할 수 있는 것이다.
  - ❏ 위 가설에 기반하여, **CLIP 이미지 인코더(마지막 layer)에 2가지 수정 적용**
    (1) 쿼리 및 키 embedding layers 제거.
    (2) 밸류 embedding layer 및 last linear layer를 1x1 conv layers로 변경.
  - ❏ 참고로, CLIP (Transformer-based) 이미지 인코더에서도
    쿼리를 클래스 토큰으로 보고 위 순리 똑같이 보면 됨.

$$\mathrm{AttnPool}(\bar{q}, k, v) = \mathcal{F}(\sum_i \mathrm{softmax}(\frac{\bar{q}k_i^\mathsf{T}}{C})v_i)$$

$$= \sum_i \mathrm{softmax}(\frac{\bar{q}k_i^\mathsf{T}}{C})\mathcal{F}(v_i), \quad (3)$$

$$\bar{q} = \mathrm{Emb_q}(\bar{x}),\ k_i = \mathrm{Emb_k}(x_i),\ v_i = \mathrm{Emb_v}(x_i), \quad (4)$$



ZHOU, Chong; LOY, Chen Change; DAI, Bo. Extract free dense labels from clip. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022. p. 696-712.
ECCV 2022 Oral

- **MaskCLIP**
  - 2가지 refinement 전략
    (1) Key smoothing
    (2) Prompt denoising
  - **Key smoothing**
    - 가설: Key features를 상응하는 각 패치의 local descriptor로 볼수있고,
      그렇다면 비슷한 key features를 갖는 패치들끼리는 비슷한 예측을 하는 것이 맞다.
    - 유사한 의미(key feature)를 가진 patch의 예측을 서로 보정
  - **Prompt denoising**
    - 전체 spatial locations에서 confidence가 전부 0.5 이하인 클래스들은 프롬프트에서 제거

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

ZHOU, Chong; LOY, Chen Change; DAI, Bo. Extract free dense labels from clip. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022. p. 696-712.

ECCV 2022 Oral

- **MaskCLIP / MaskCLIP+**

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.

ZHOU, Chong; LOY, Chen Change; DAI, Bo. Extract free dense labels from clip. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022. p. 696-712.

- **Experiments**

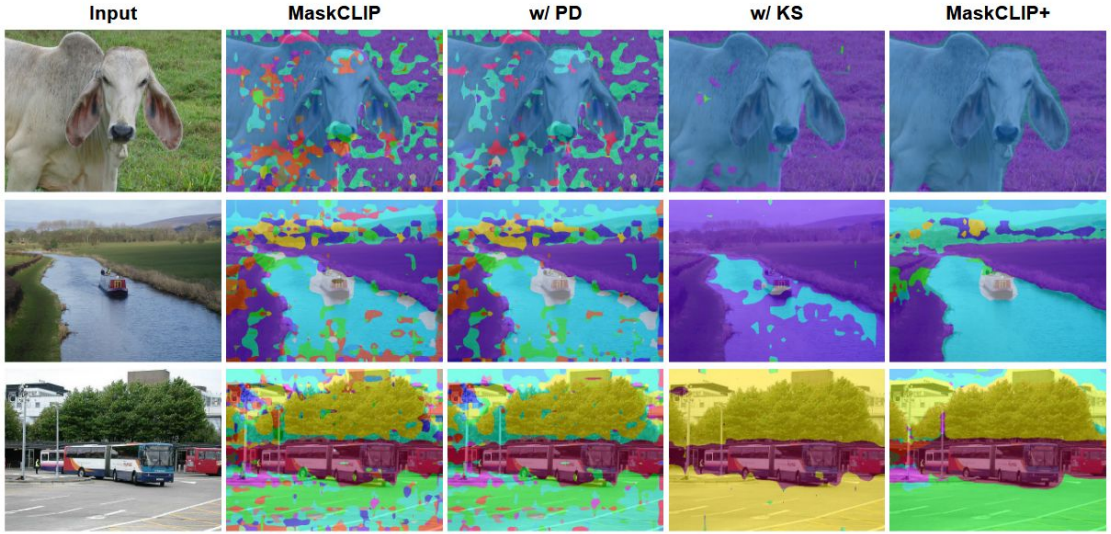


Fig. 3: **Qualitative results on PASCAL Context.** Here all results are obtained **without** any annotation. PD and KS refer to prompt denoising and key smoothing respectively. With PD, we can see some distraction classes are removed. KS is more aggressive. Its outputs are much less noisy but are dominated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

ZHOU, Chong; LOY, Chen Change; DAI, Bo. Extract free dense labels from clip. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022. p. 696-712.

ECCV 2022 Oral

- **Experiments**

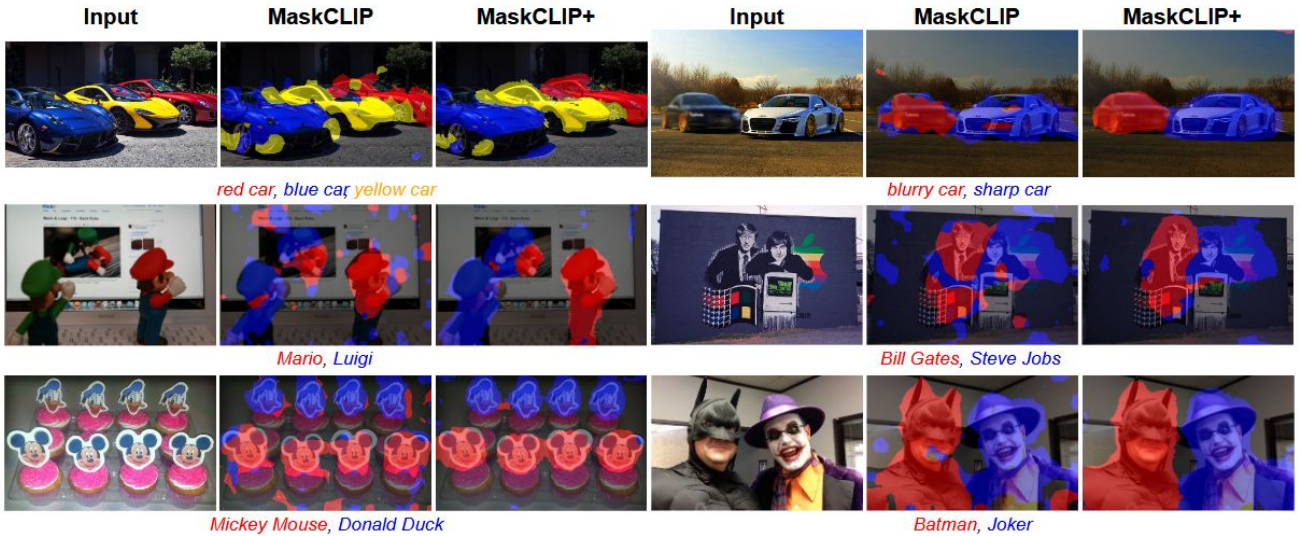


Fig. 4: **Qualitative results on Web images.** Here we show the segmentation results of MaskCLIP and MaskCLIP+ on various **unseen classes**, including fine-grained classes such as cars in different colors/imagery properties, celebrities, and animation characters. All results are obtained **without**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