ECCV 2024

ProxyCLIP: Proxy Attention Improves CLIP for Open-Vocabulary Segmentation

Mengcheng Lan[1], Chaofeng Chen[1], Yiping Ke[2], Xinjiang Wang[3], Litong Feng[3], and Wayne Zhang[3,4*]

[1] S-Lab, Nanyang Technological University
[2] CCDS, Nanyang Technological University [3] SenseTime Research
[4] Guangdong Provincial Key Laboratory of Digital Grid Technology
lanm0002@e.ntu.edu.sg {chaofeng.chen, ypke}@ntu.edu.sg
{wangxinjiang, fenglitong, wayne.zhang}@sensetime.com
https://github.com/mc-lan/ProxyCLIP

- Problem / objective
  - CLIP 사용해서 Open-Vocabulary Semantic Segmentation
    - Recognition 잘하지만, localization을 못함
- Contribution / Key idea
  - **ProxyCLIP**: OVSS using "VFM의 robust local consistency + CLIP의 zero-shot transfer capacity"
    - Localization 문제를 VFM의 spatial feature correspondence으로 보완
    - CLIP의 마지막 layer에 Proxy Attention Module(PAM) 적용

전유진

- **Motivation**

  - CLIP: Recognition 굿, but localization 별로
  - VFM: Semantic understanding 별로, but spatial coherence 굿
  - **ProxyCLIP:** CLIP의 단점을 VFM의 장점으로 극복
    - VFM의 feature correspondence를 CLIP에 proxy attention을 통해 사용
    - VFM의 feature correspondence 사용할때 2가지 전략
      - (1) Adaptive normalization
      - (2) Masking strategy

- **Motivation**

  - ❑ CLIP의 attention map $\mathrm{Attn}_{qk} \in \mathbb{R}^{L \times L}$
  - ❑ VFM의 feature correspondence map $F \in \mathbb{R}^{L_v \times D_v}$ $S_{ij} = \frac{F_i}{|F_i|}\frac{F_j}{|F_j|}$
  - ❑ Semantic coherence
    CLIP의 vanilla attention (q-k) < self-self attention (q-q/k-k/v-v) < VFM의 feature correspondence

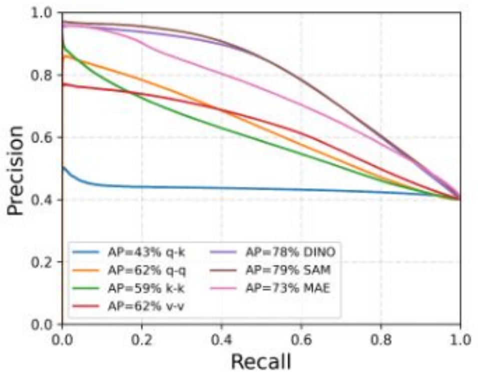


Fig. 1: Precision recall curves of different classifiers. Higher average precision (AP) indicates better semantic correspondence.

LAN, Mengcheng, et al. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024. p. 70-88.

- **Motivation**
  - ❑ Semantic coherence
    CLIP의 vanilla attention (q-k) < self-self attention (q-q/k-k/v-v) < VFM의 feature correspondence
  - ❑ 목표
    VFM의 advanced spatial coherence와 CLIP의 semantic understanding 능력을 training-free framework 하에 결합하여, vision-language inference 성능을 향상시키겠다.

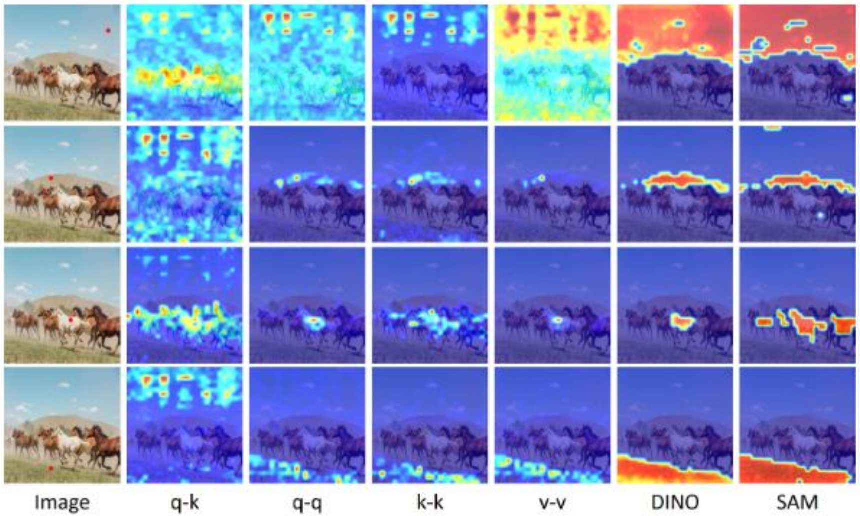


Fig. 2: Attention scores (maps) between CLIP, DINO and SAM using different seeds (in red). For CLIP's attention maps, we display only the first head of multi-head self-attention maps.

전유진

- **Overview**

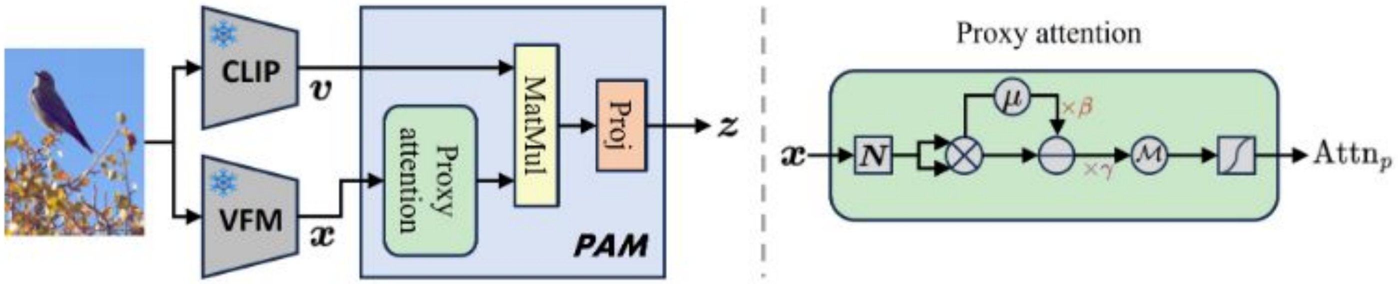


Fig. 3: Overview of the ProxyCLIP architecture. ProxyCLIP consists of two frozen image encoders and a novel proxy attention module (PAM). On the right, the flow of the proxy attention mechanism with an adaptive normalization and masking strategy is illustrated, corresponding to Eqs. (6) to (8).

- **ProxyCLIP**

❑ Proxy Attention Module (PAM)
- 쿼리, 키: VFM features, 밸류: CLIP features
- 주의) Cross-attention 아니고, Self-attention 기반 Proxy 구조임.

$$\mathrm{Attn}_p = \mathrm{SoftMax}(\boldsymbol{x}\boldsymbol{x}^T), \quad (4)$$

$$\boldsymbol{z} = \mathrm{Proj}(\mathrm{Attn}_p \cdot \boldsymbol{v}), \quad (5)$$

❑ Normalization and Masking
- 문제: 수식4에 기반한 proxy attention score가 다양한 VFM 간에 항상 좋은 일관성과 분리성을 보장하지 않는다.
- 원인: 각 VFM의 visual representations의 서로 다른 inductive biases에 기인.

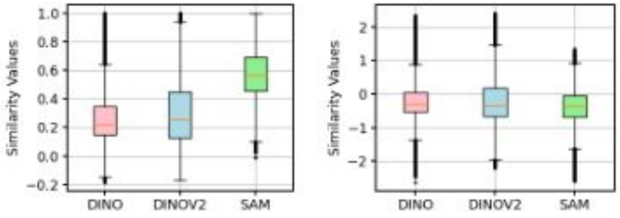


Fig. 4: The statistics of similarity matrix before (left) and after (right) normalization.

- **ProxyCLIP**

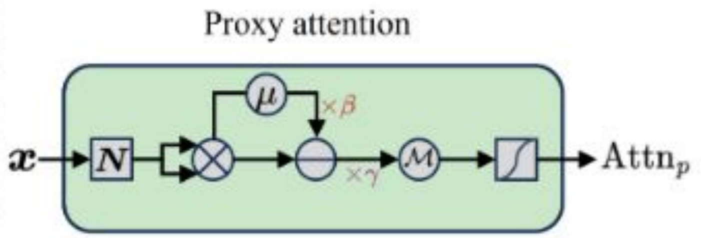
❑ Normalization and Masking
- 문제: 수식4에 기반한 proxy attention score가 다양한 VFM 간에 항상 좋은 일관성과 분리성을 보장하지 않는다.
- 원인: 각 VFM의 visual representations의 서로 다른 inductive biases에 기인.
- 해결: Normalization and Masking
  Masking matrix M가 Normalized similarity matrix A에서 negative similarities에 해당하는 패치들을 suppress함.
  (베타값 1.2, 감마값 3)

❑ Different resolution
공간 해상도 높게 (패치 개수 많게) proxy attention하고, interpolation을 통해 x와 v의 spatial resolution 맞춤.

$$A = \gamma(\boldsymbol{x}\boldsymbol{x}^T - \frac{\beta}{L_v^2}\sum_{i,j}[\boldsymbol{x}\boldsymbol{x}^T]_{ij}), \quad (6)$$

$$\mathcal{M}_{ij} = \begin{cases} 0, & A_{ij} \geq 0 \\ -\infty, & A_{ij} < 0 \end{cases} \quad (7)$$

$$\mathrm{Attn}_p = \mathrm{SoftMax}(A + \mathcal{M}). \quad (8)$$



LAN, Mengcheng, et al. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024. p. 70-88.

- **Experiments**

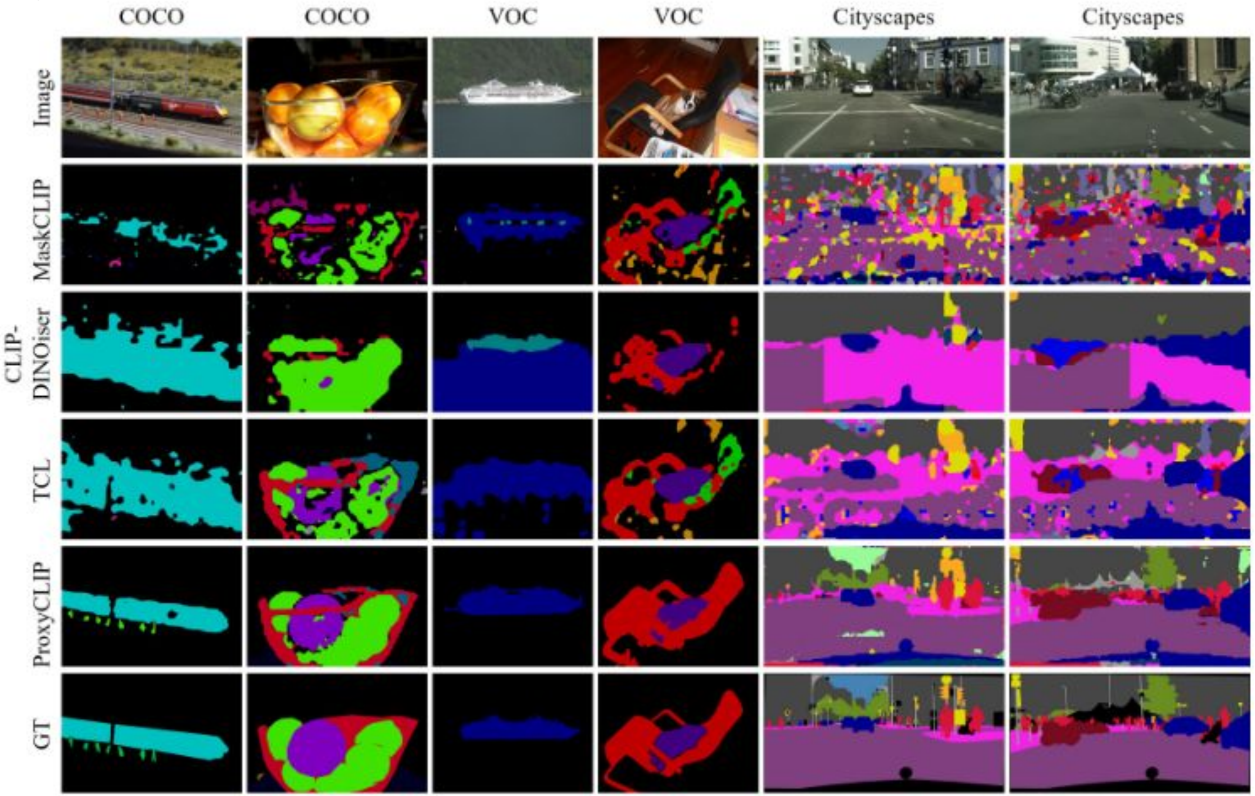


Fig. 5: Qualitative comparison of semantic segmentation results.

- **Experiments**

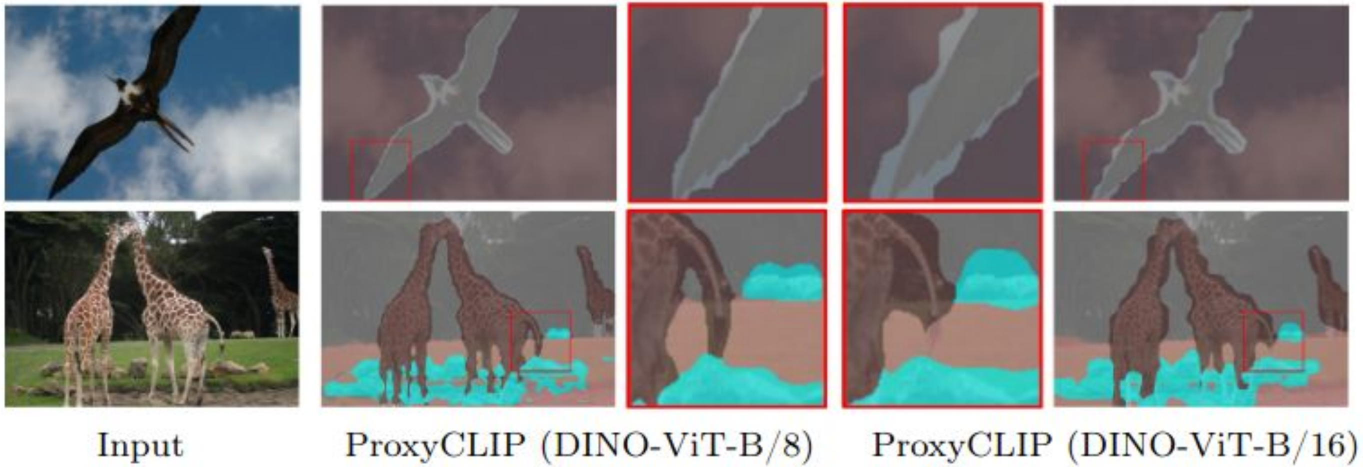


Fig. 7: Qualitative comparison of different patch size. VFMs with smaller patch size of 8 helps ProxyCLIP to produce sharper boundaries.