MasQCLIP for Open-Vocabulary Universal Image Segmentation

Xin Xu[1*+], Tianyi Xiong[2*], Zheng Ding[3], Zhuowen Tu[3]
[1]Peking University, [2]Tsinghua University, [3]University of California, San Diego
*Indicate equal contribution. +Work done during internship at University of California, San Diego.

Acknowledgement: *This work is supported by NSF Award IIS-2127544.*

ICCV23

Project Page: https://masqclip.github.io/

Introduction

(a) "a black cat", "an orange cat" (b) "beer", "wine" (c) "an adult", "a child"

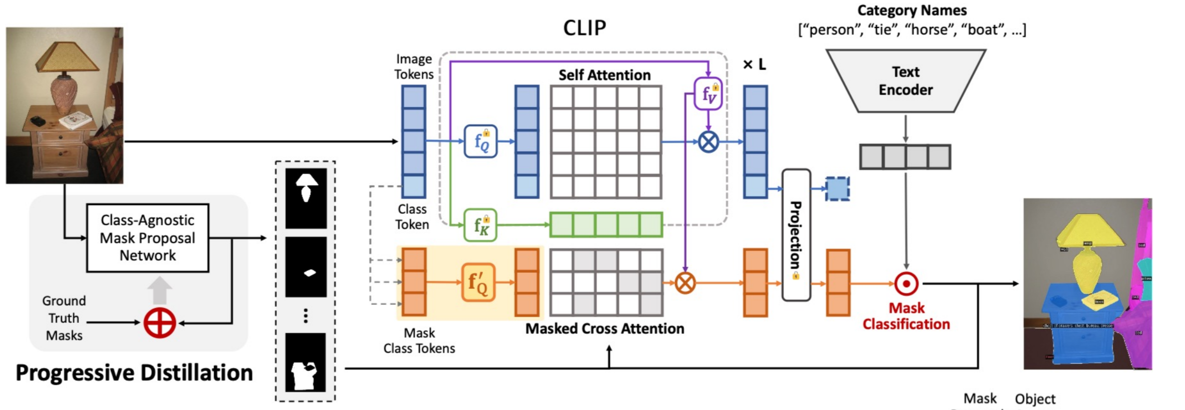
- **Being Open-Vocabulary:** Target of interest to be extracted can be freely specified using natural language description during inference.
- **Being Universal:** Perform instance, semantic, and panoptic segmentation under a unified framework.

Motivation

- CLIP aligns images and texts into the same feature space but **cannot discriminate between objects of the same category**.
- Previous works have difficulty in generating new mask proposals beyond supervision; and lack in adaption to mask classification due to the gap between image-level and region-level representation.

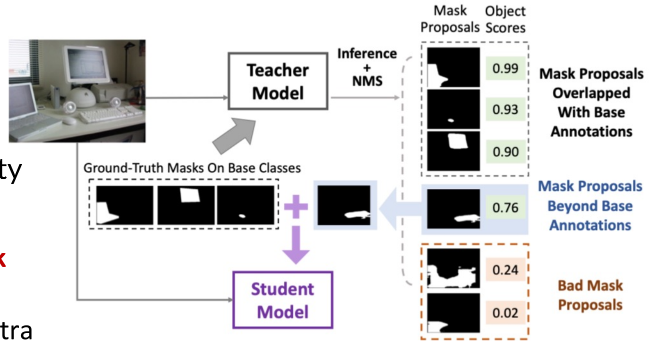
How to balance between **maintaining generalization for more categories** and **adapting CLIP for mask classification?**

Method



Progressive Distillation (stage 1)

- CLIP does not intrinsically assign higher confidence-scores to good-quality masks.
- **Object Score**: general indicator of mask quality
  Final Classification Score: $p_{cls}^{(i)} = p_{obj} \cdot p_{clip}^{(i)}$
- Utilize object score to **filter high-quality mask proposals that do not overlap with mask annotations of base categories**, producing extra annotations for training.



Detailed Interpretation of Progressive Distillation

MasQ-Tuning (stage 2)

- For i-th Mask Class Token $x_{mask}^{(i)}$ and its query embedding $q_i = f_Q(x_{mask}^{(i)})$, its attention weight $\mathrm{softmax}(q_i K_{img}^T + M_i)$ indicates $x_{mask}^{(i)}$ where to focus.
- We **apply new query projections f_Q'** to each cross-attention layer for Mask Class Tokens but **keep the original CLIP frozen**
  $\mathrm{CrossAttn}(\cdot) = \mathrm{softmax}\left(\mathbf{Q'_{mask}} \mathrm{K}_{\mathrm{img}}^{\mathrm{T}} + \mathcal{M}_{mask}\right) \cdot V_{img}$
  where $\mathbf{Q'_{mask}} = f_Q'(x_{mask})$.
- Mask Class Tokens obtain better attention weights through learning while the cross-attention results still lie in the row space of V_{img}. **Able to improve adaptation (from image to mask classification) while maintaining generalization.**

Preliminary: MaskCLIP[1]

Mask Class Tokens extract features from CLIP tokens through masked cross-attention mechanism where mask proposals serve as attention masks.

$$\mathrm{CrossAttn}(\cdot) = \mathrm{softmax}(Q_{\mathrm{mask}} K_{\mathrm{img}}^T + \mathcal{M}_{\mathrm{mask}}) \cdot V_{\mathrm{img}}$$

$$Q_{\mathrm{mask}}, K_{\mathrm{img}}, V_{\mathrm{img}} = f_Q(x_{\mathrm{mask}}), f_K(x_{\mathrm{img}}), f_V(x_{\mathrm{img}})$$

$$\mathcal{M}_{\mathrm{mask}}(i,j) = \begin{cases} 0 & \text{if } i\text{-th mask falls in } j\text{-th patch} \\ -\infty & \text{otherwise} \end{cases}$$

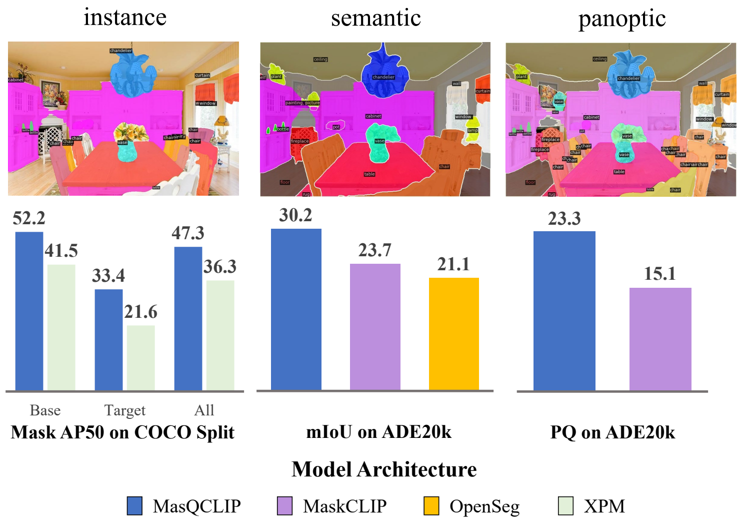
[1] Z. Ding, J. Wang, and Z. Tu. Open-Vocabulary Universal Image Segmentation with MaskCLIP. In *ICML*, 2023

Quantitative Results

Methods	Instance			Semantic				Panoptic		
	Base	Novel	All	A-150	A-847	P-59	P-459	PQ	PQ	PQ
XPM	41.5	21.6	36.3	-	-	-	-	-	-	-
LSeg+	-	-	-	18.0	3.8	46.5	7.8	-	-	-
OpenSeg	-	-	-	21.1	6.3	42.1	9.0	-	-	-
MaskCLIP	-	-	-	23.7	8.2	45.9	10.0	15.1	13.5	18.3
MasQCLIP	**51.0**	**31.9**	**46.0**	**30.4**	**10.7**	**57.8**	**18.2**	**23.3**	**21.2**	**27.7**
	+9.5	+10.3	+9.7	+6.7	+2.5	+11.3	+8.2	+8.2	+7.7	+9.4

Achieve **substantial performance gain** across all open-vocabulary segmentation tasks **with a unified framework**.

Visualization



Panoptic Segmentation: MasQCLIP is able to segment both thing(object) and stuff(background) categories more correctly.