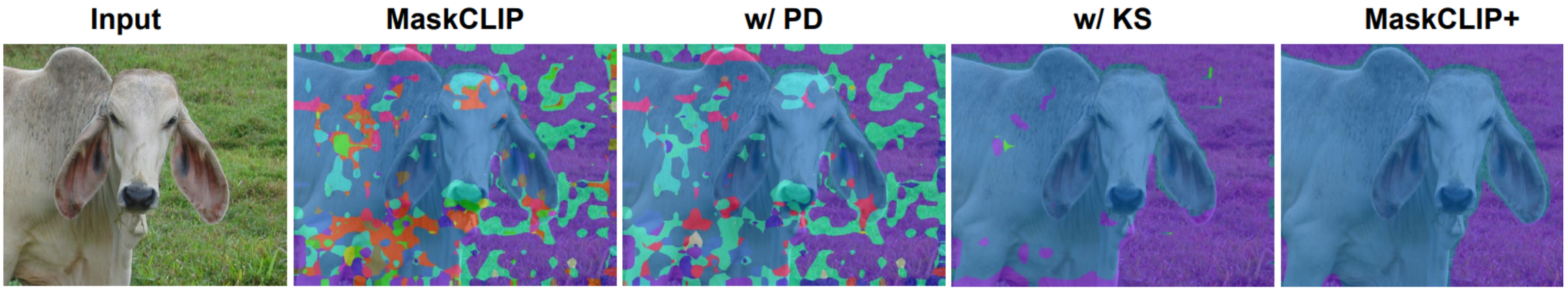
Extract Free Dense Labels from CLIP

Authors: Chong Zhou, Chen Change Loy , and Bo Dai



Sonia Castro Paniello

20.02.24

INTRODUCTION

Goal: Dense prediction with CLIP

CLIP

Learning Transferable Visual Models From Natural Language Supervision

→ **Dense Prediction ?**

Goal: Dense prediction with CLIP

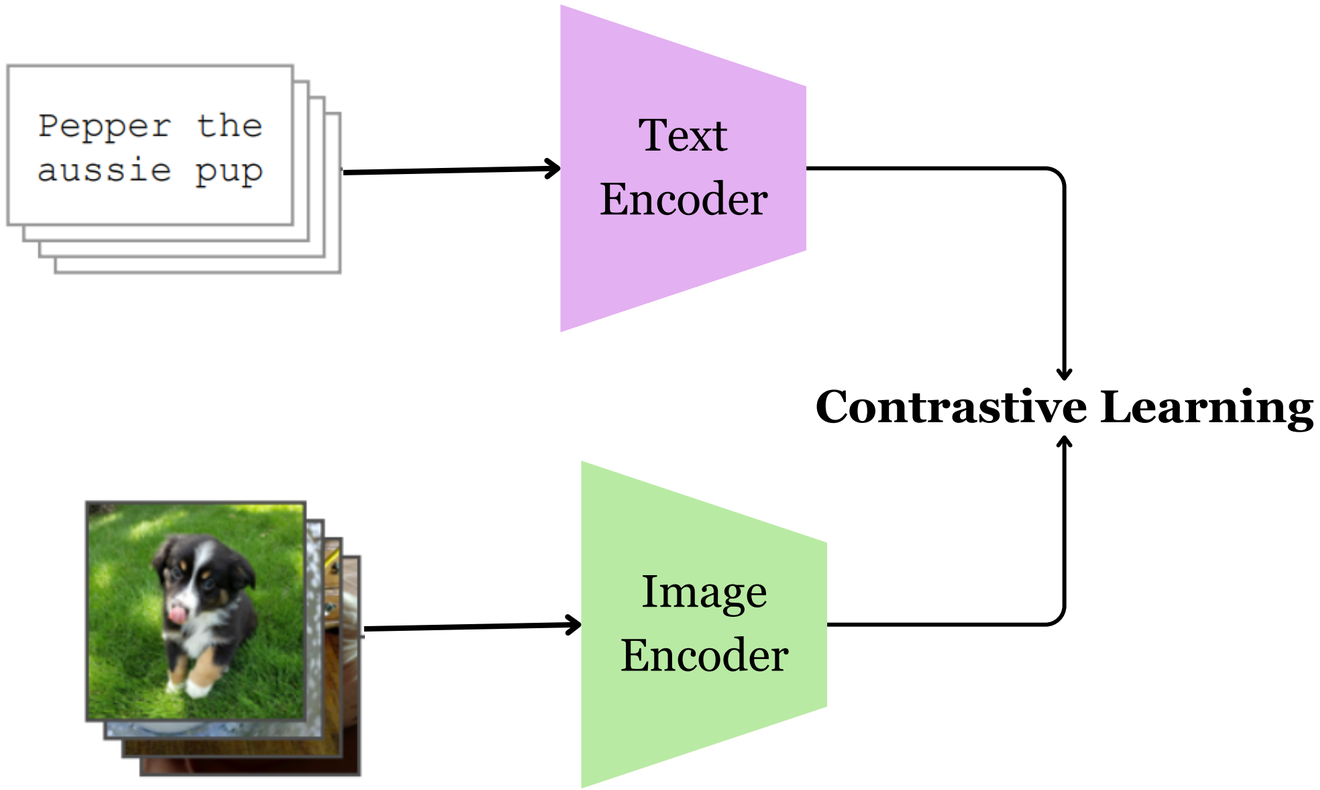
CLIP

Learning Transferable Visual Models From Natural Language Supervision

Dense Prediction ?

- Expensive Annotations
- Open Vocabulary

CLIP: Contrastative Image-Language Pretraining



CLIP: Contrastative Image-Language Pretraining

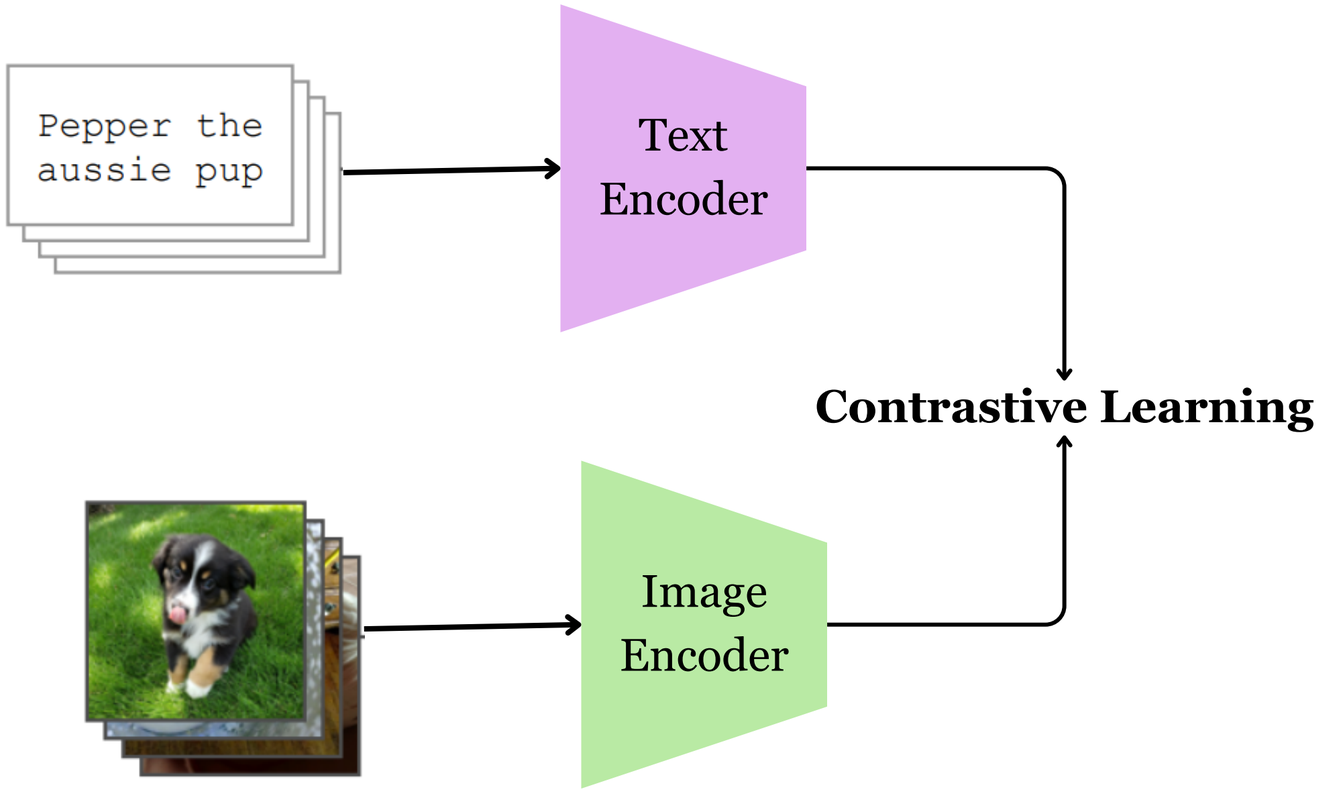


Image 1	Text 1
Image 2	Text 2
Image 3	Text 3
Image 4	Text 4
.	.
.	.
.	.
.	.
.	.
Image N	Text N

CLIP: Contrastative Image-Language Pretraining

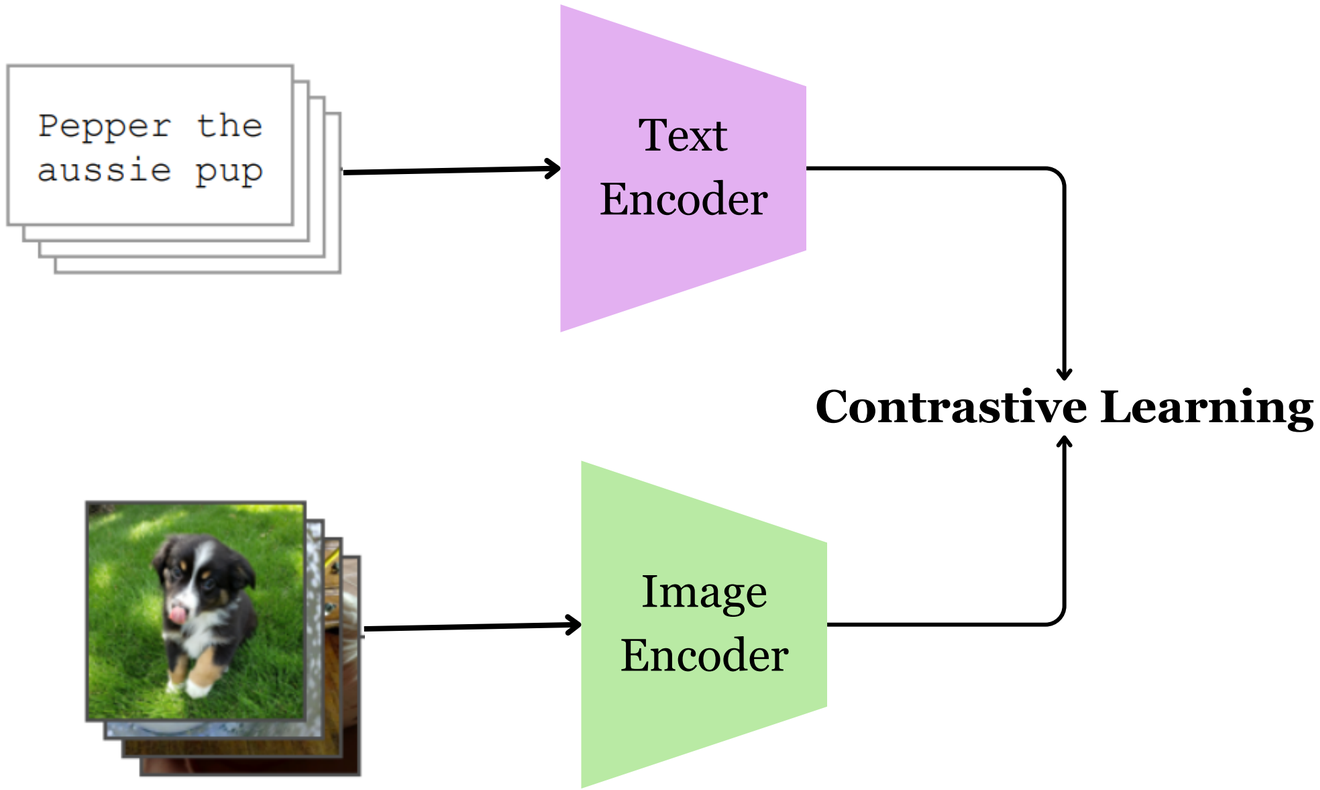


Image 1	Text 1
Image 2	Text 2
Image 3	Text 3
Image 4	Text 4
.	.
.	.
.	.
.	.
.	.
Image N	Text N

CLIP: Contrastative Image-Language Pretraining

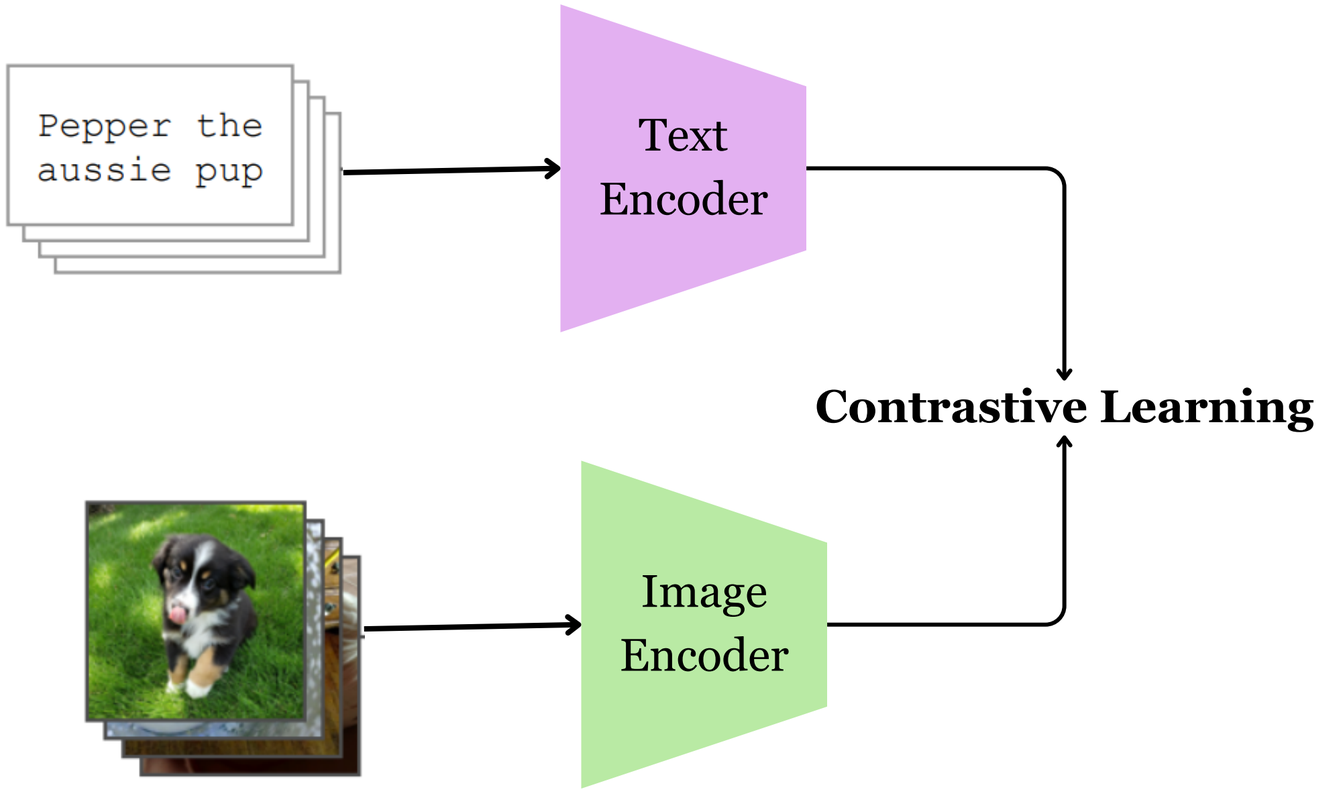


Image 1	Text 1
Image 2	Text 2
Image 3	Text 3
Image 4	Text 4
.	.
.	.
.	.
.	.
.	.
Image N	Text N

CLIP: Contrastative Image-Language Pretraining

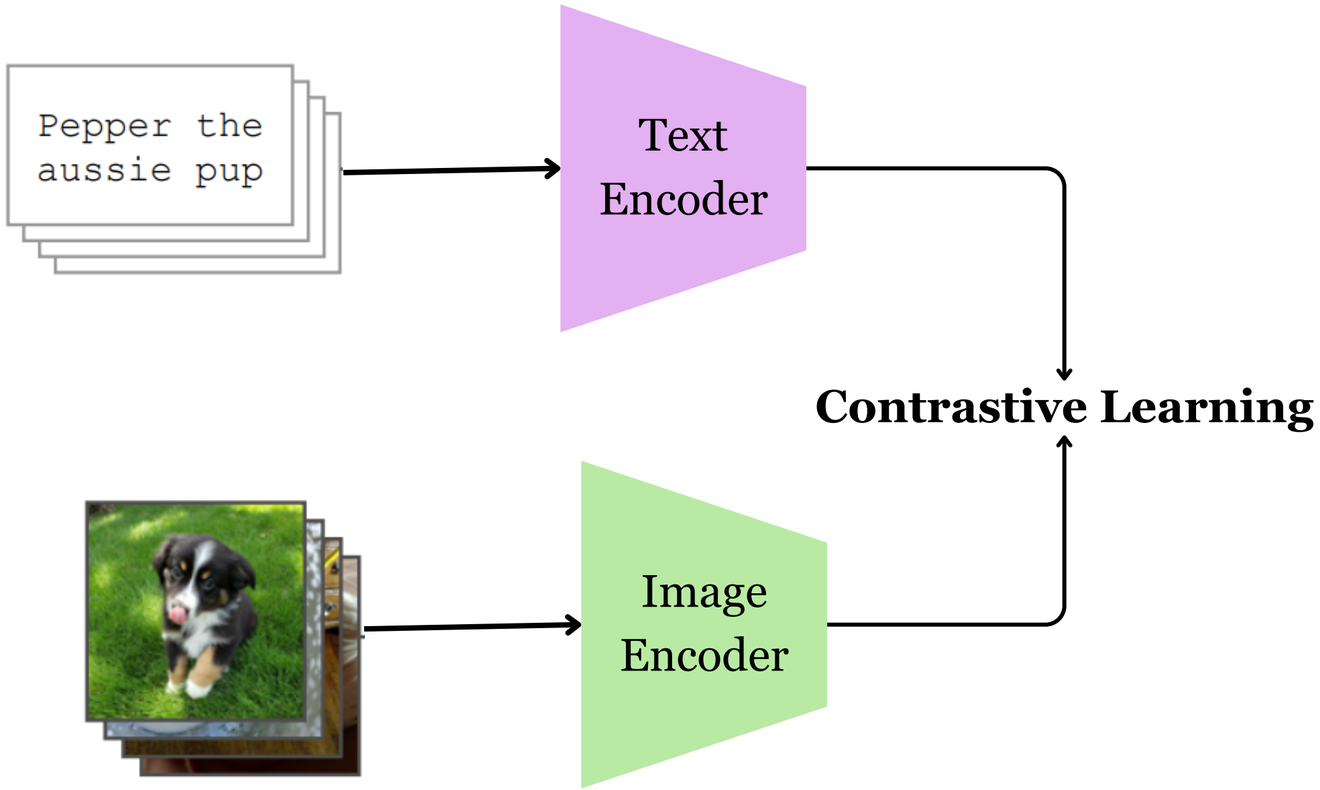
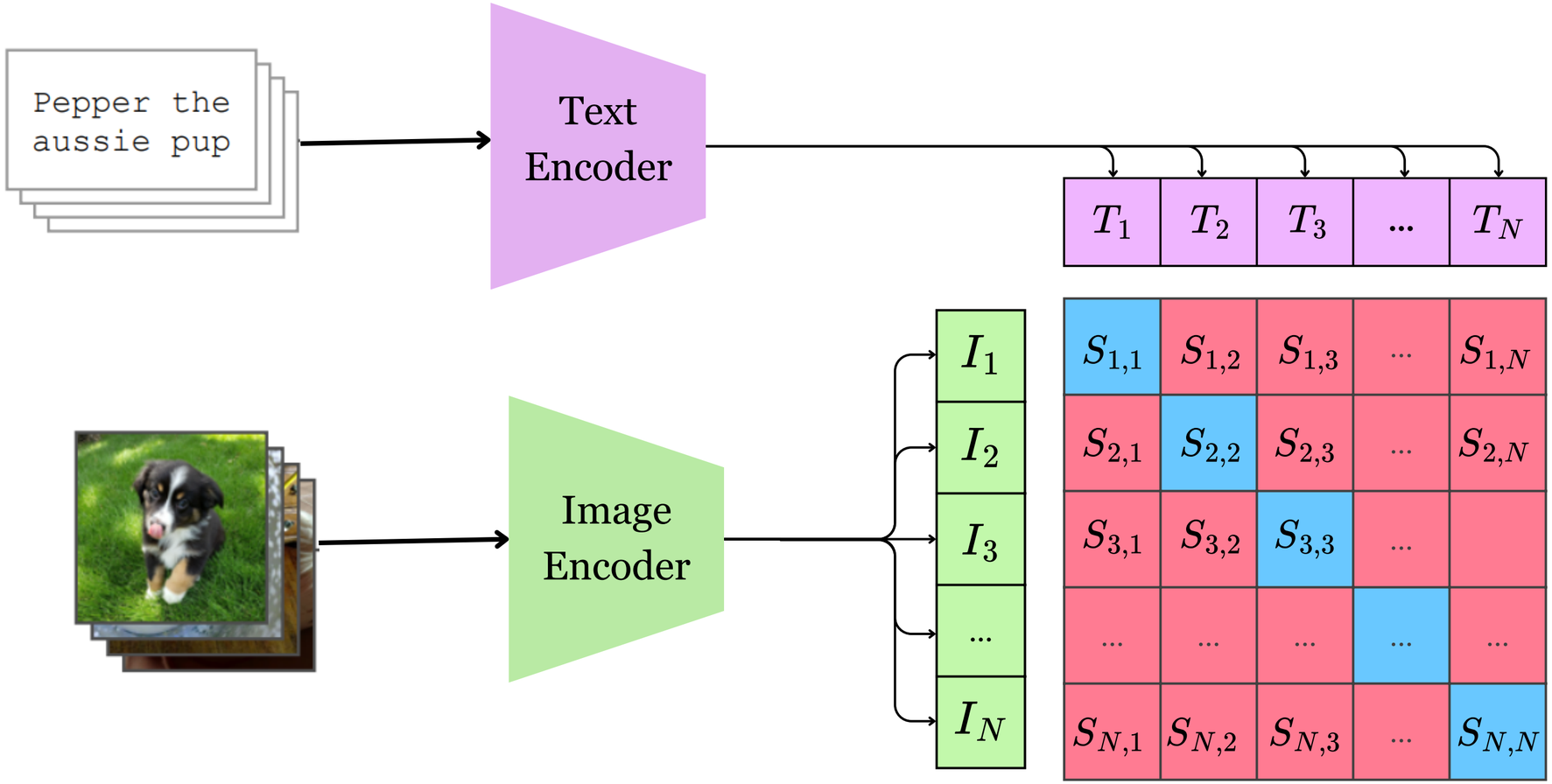


Image 1	Text 1
Image 2	Text 2
Image 3	Text 3
Image 4	Text 4
.	.
.	.
.	.
.	.
.	.
Image N	Text N

CLIP: Contrastative Image-Language Pretraining



Intuition: Descriptions contain Dense Semantic Guidance

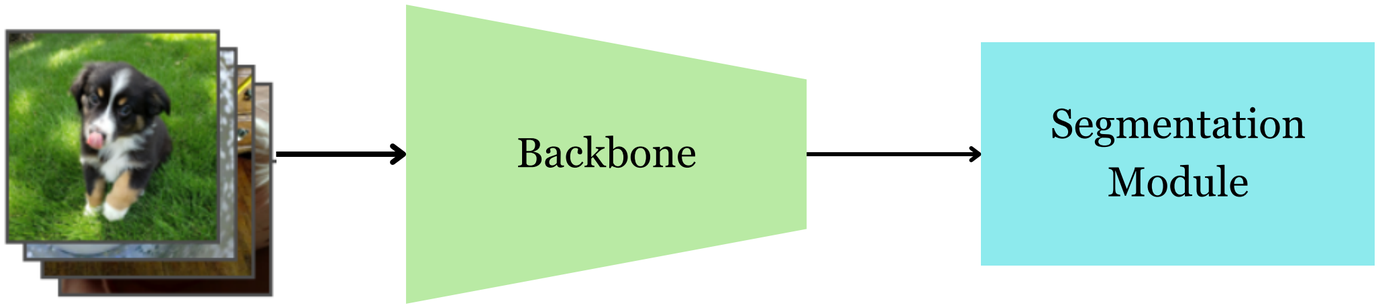
The man at bat readies to swing at the pitch while the umpire looks on.

A horse carrying a large load of hay and two people sitting on it.

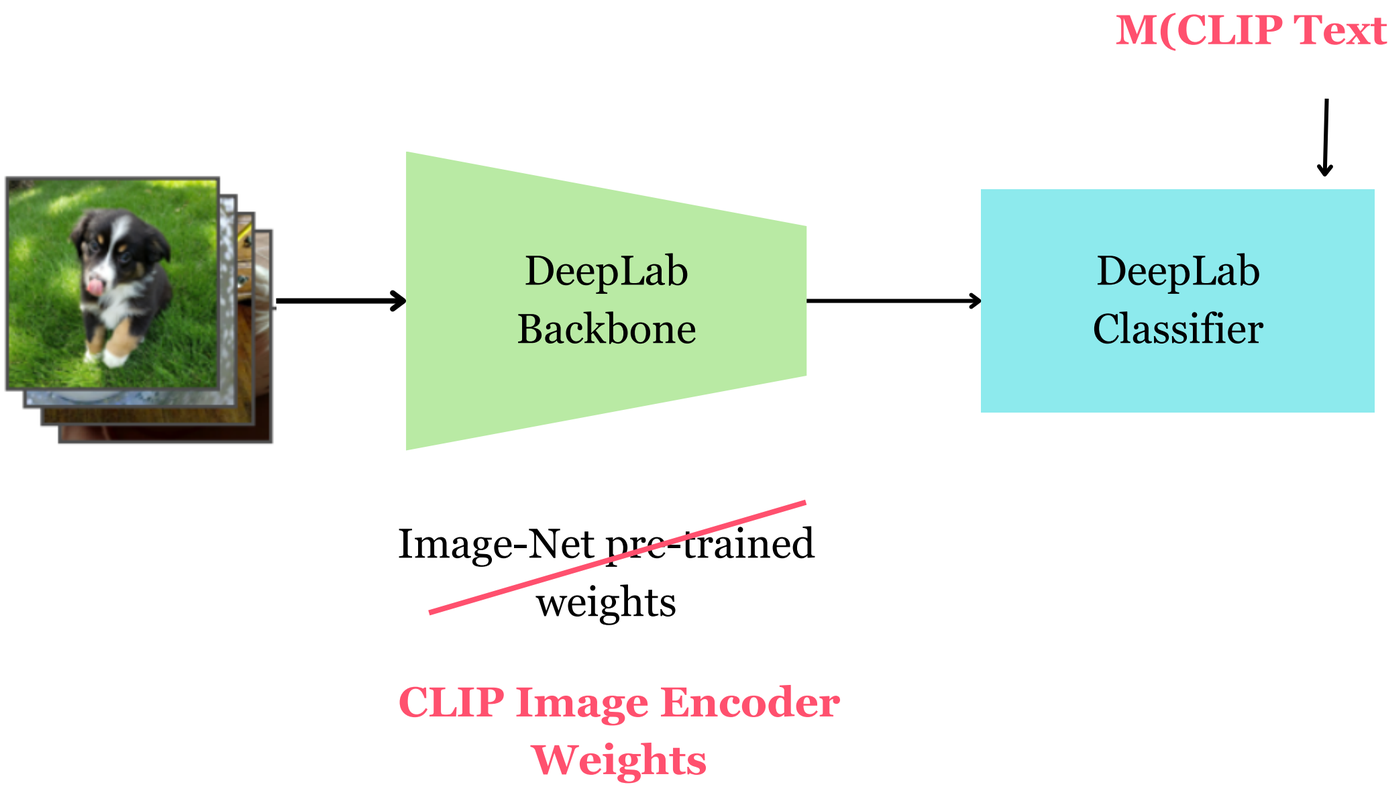
- Local image semantics
- Open vocabulary
- Contextual information

METHOD

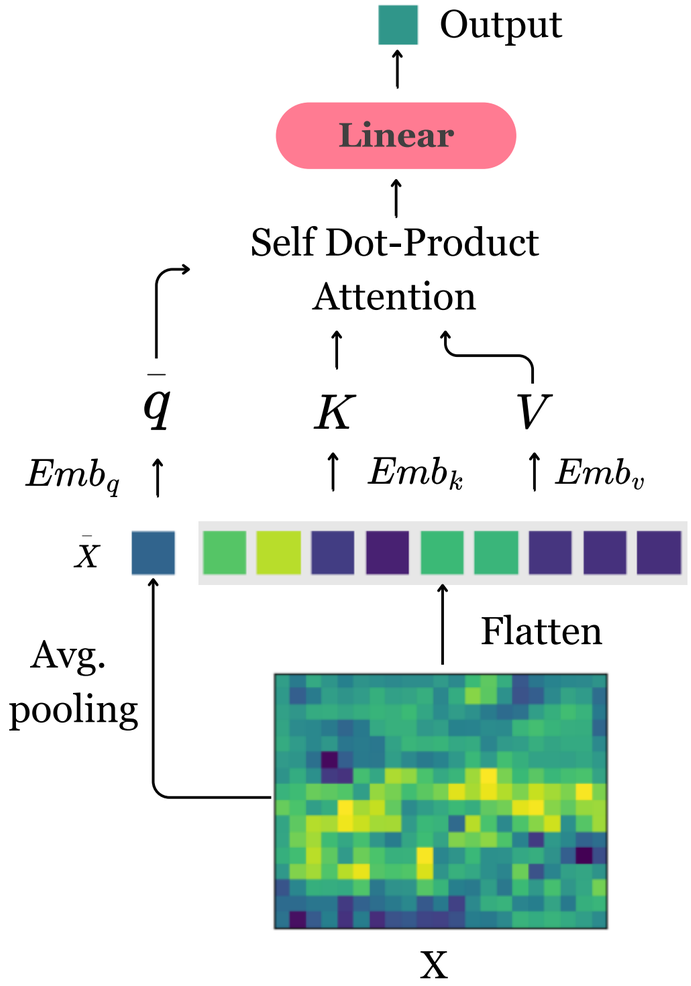
First Approach: Backbone + Segmentation Module



First Approach: Fails to segment unannotated classes

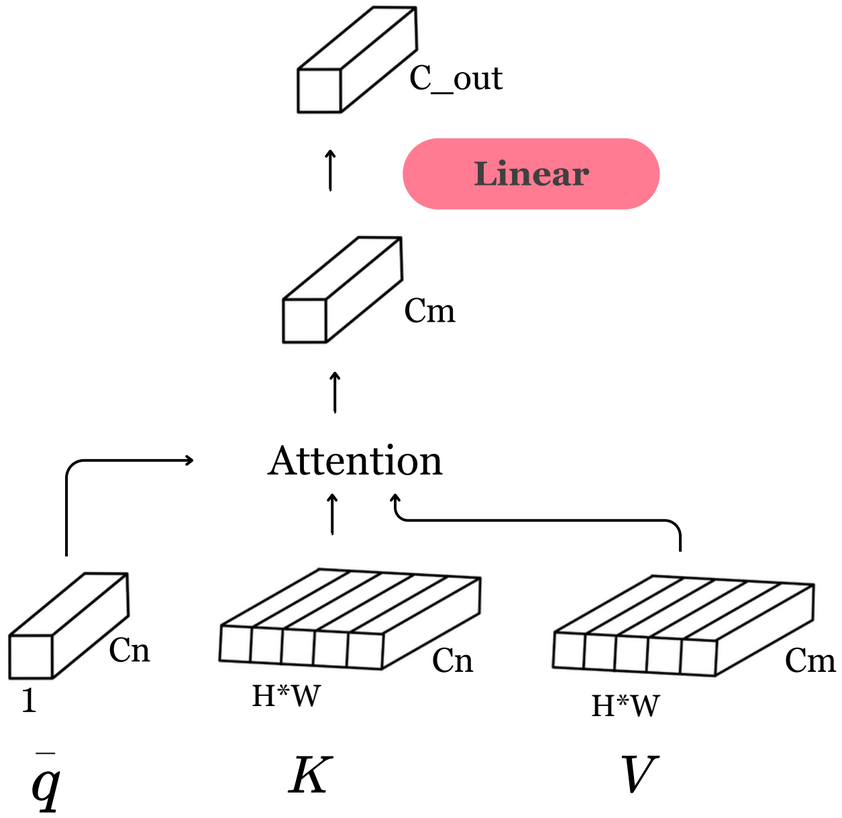


CLIP: Global Attention Pooling

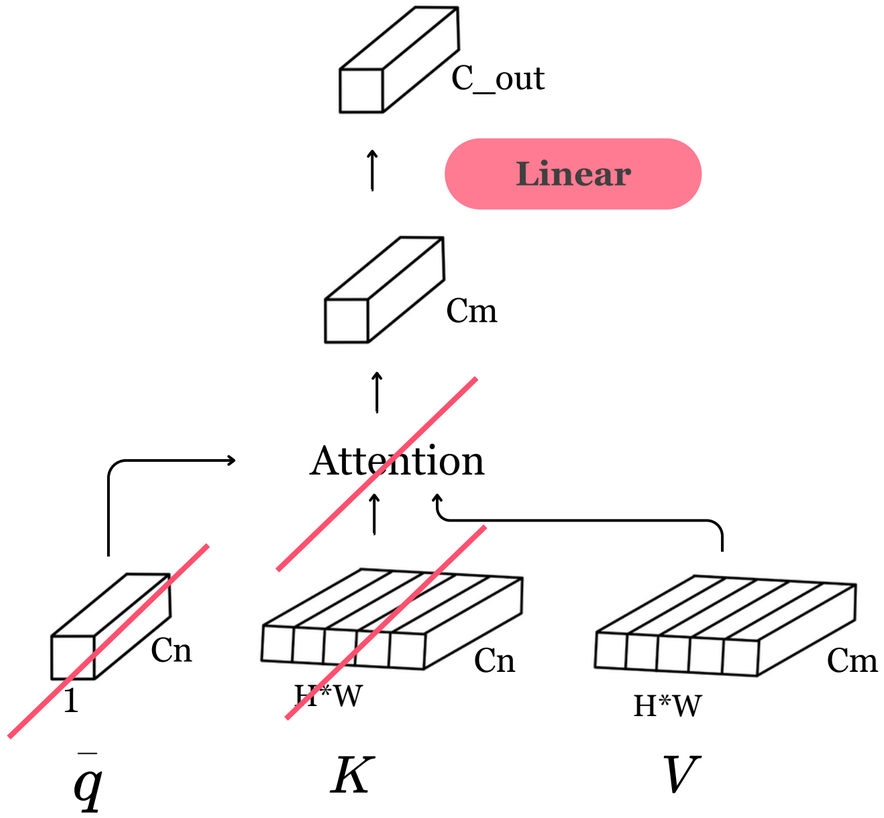


$$\begin{aligned}\text{AttnPool}(\bar{q}, k, v) &= \mathcal{F}(\sum_i \text{softmax}(\frac{\bar{q}k_i^\intercal}{C})v_i) \\ &= \sum_i \text{softmax}(\frac{\bar{q}k_i^\intercal}{C})\mathcal{F}(v_i),\end{aligned}$$

CLIP: Global Attention Pooling

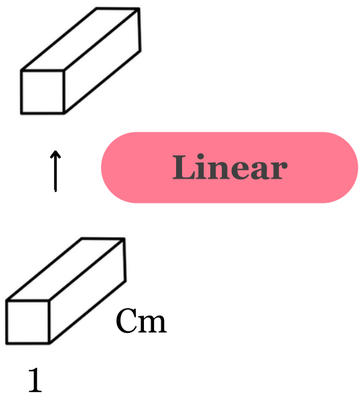


Remove Query and Key Embeding Layers



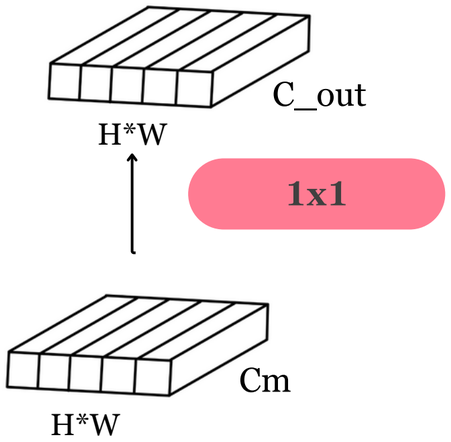
Reformulate Linear Layer to 1x1 Convolution

CLIP



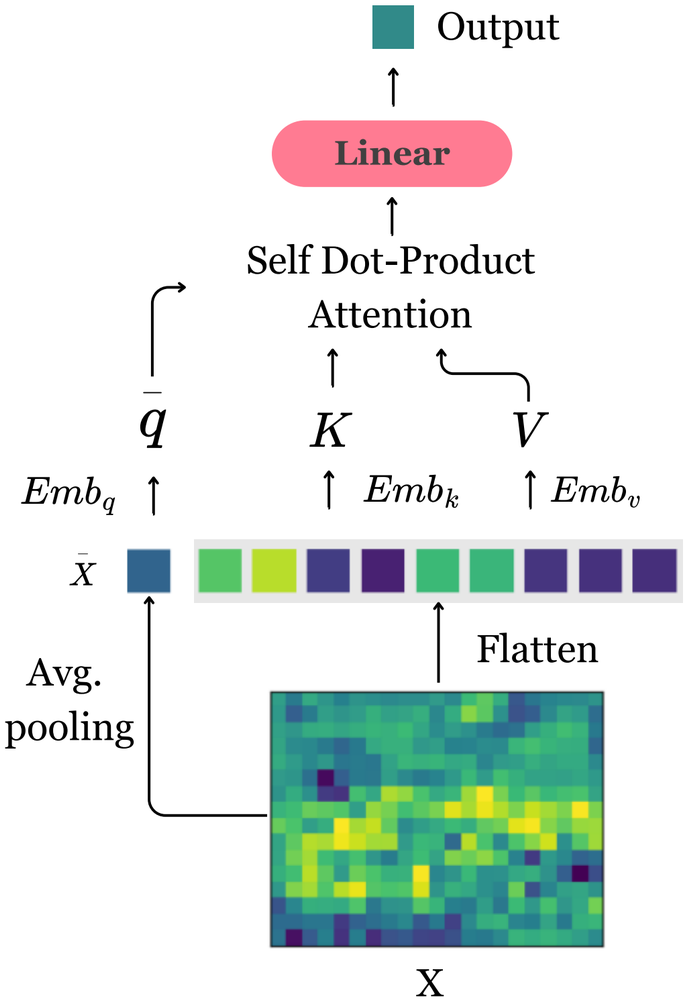
Attention Output

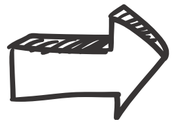
MASKCLIP

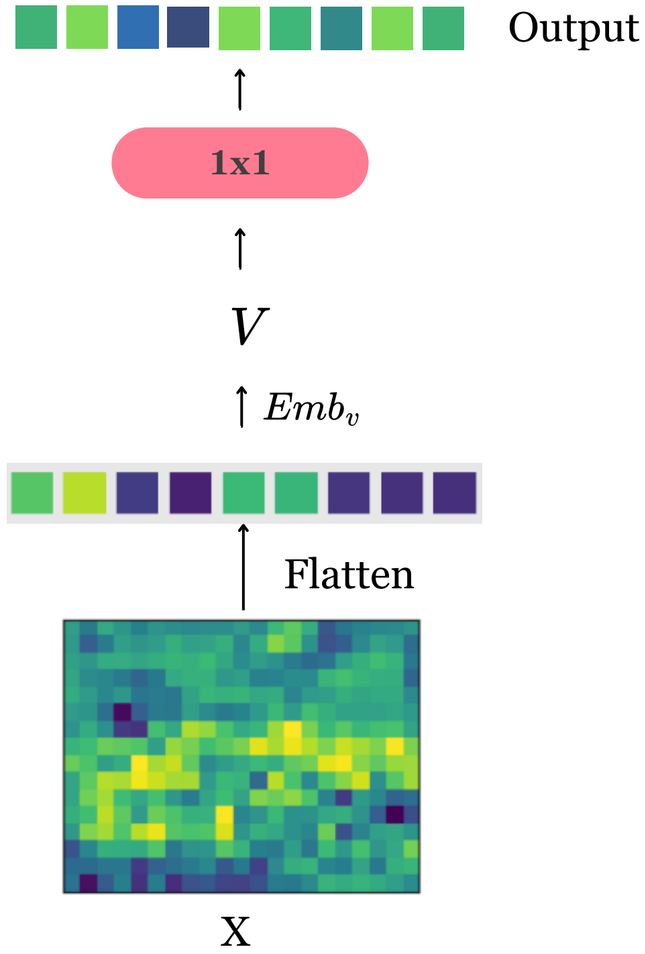


V

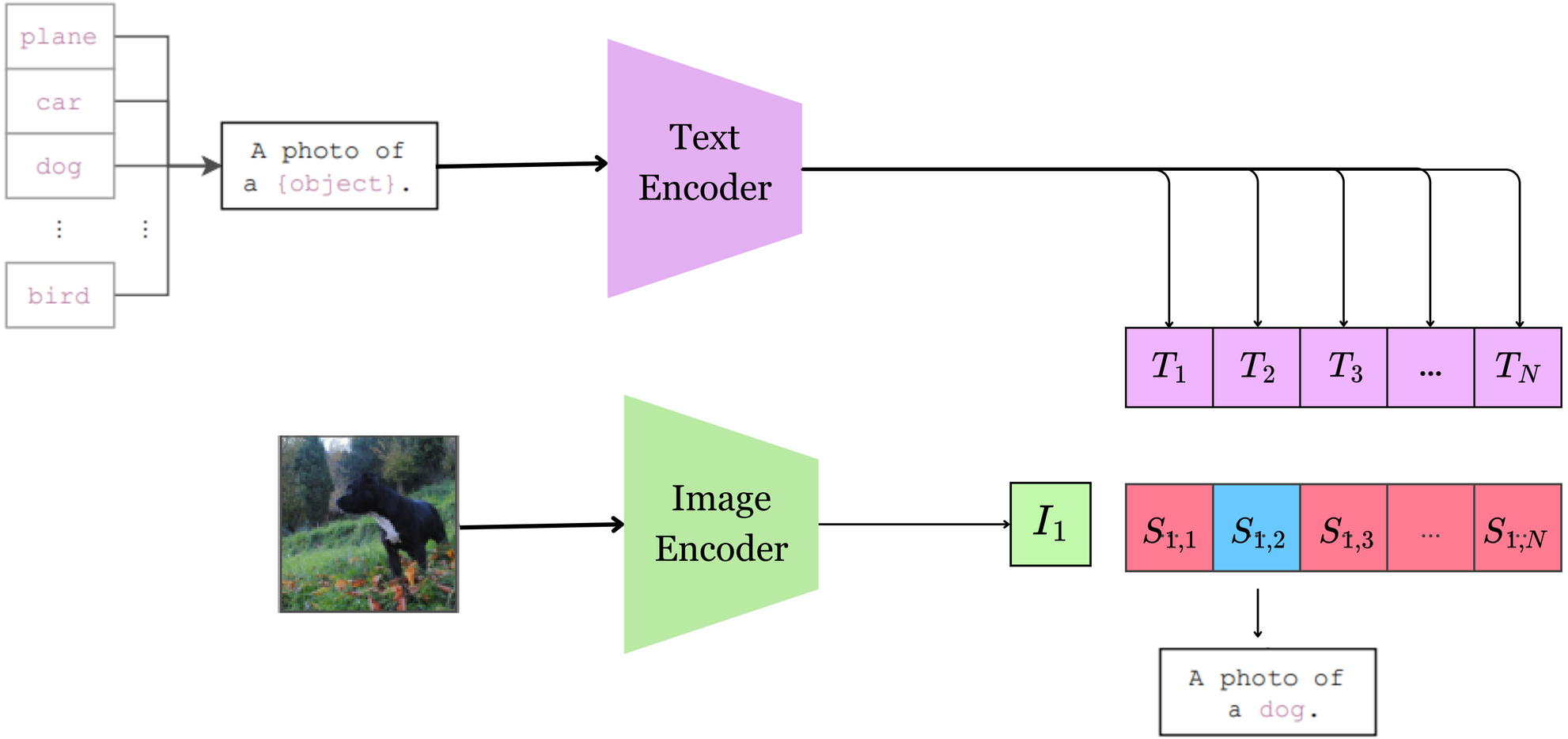
MaskCLIP: Image Encoder



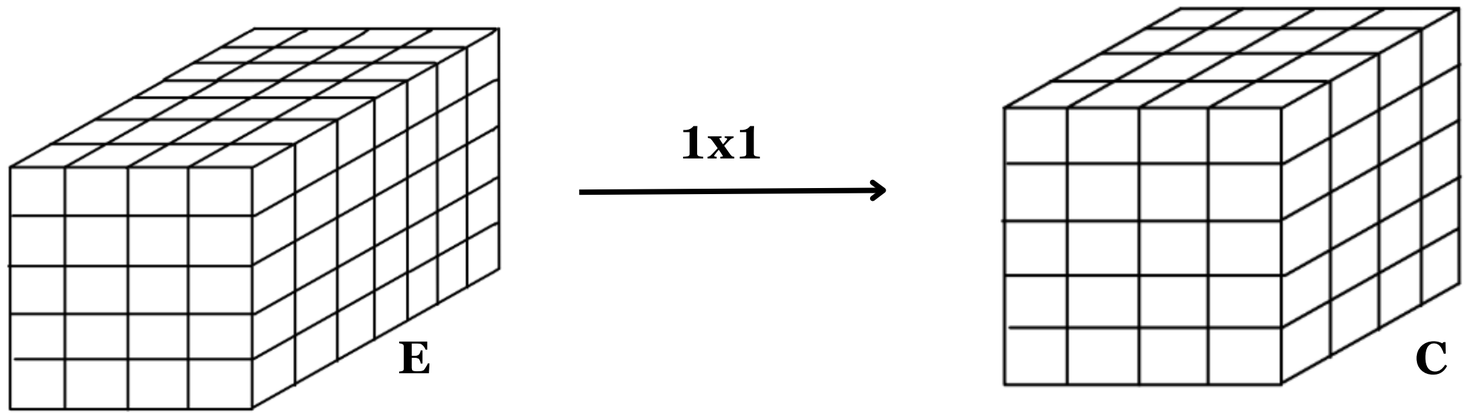




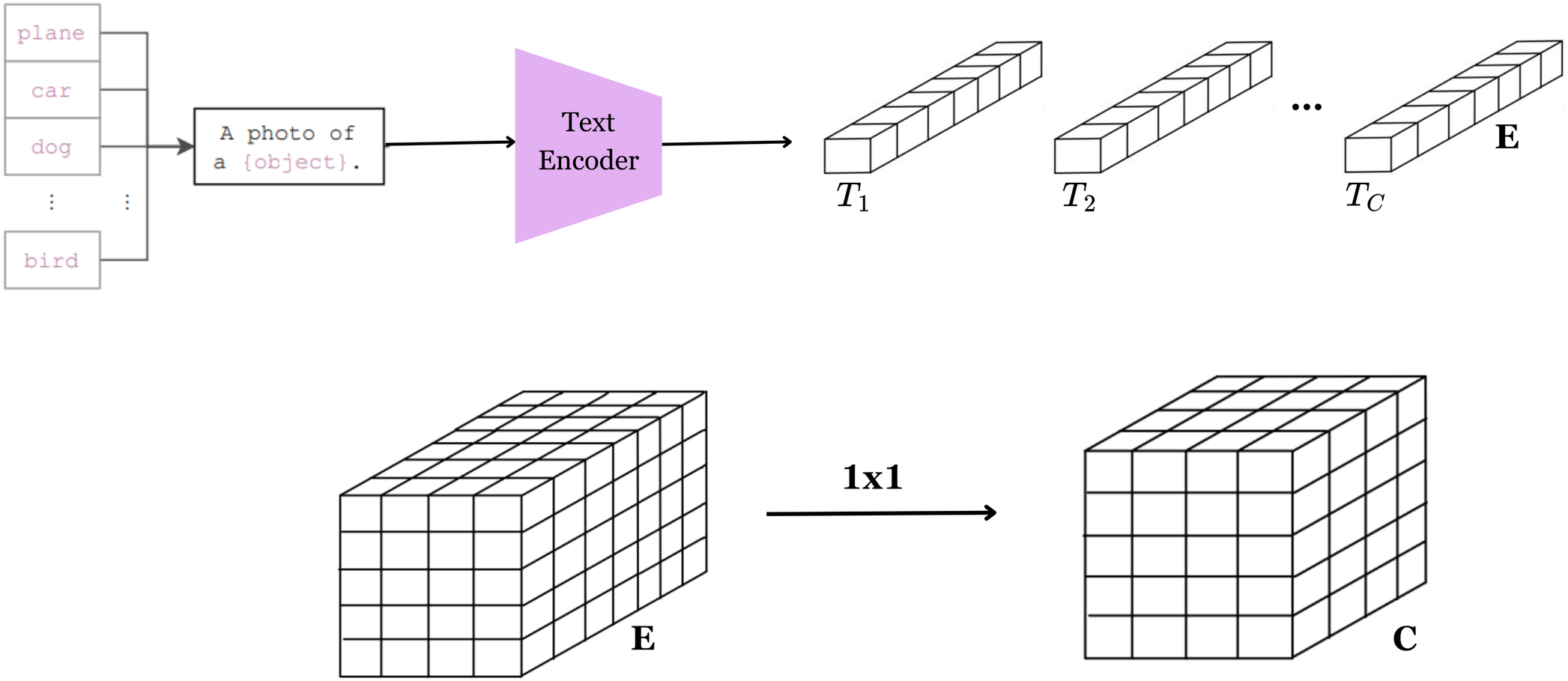
CLIP: Inference



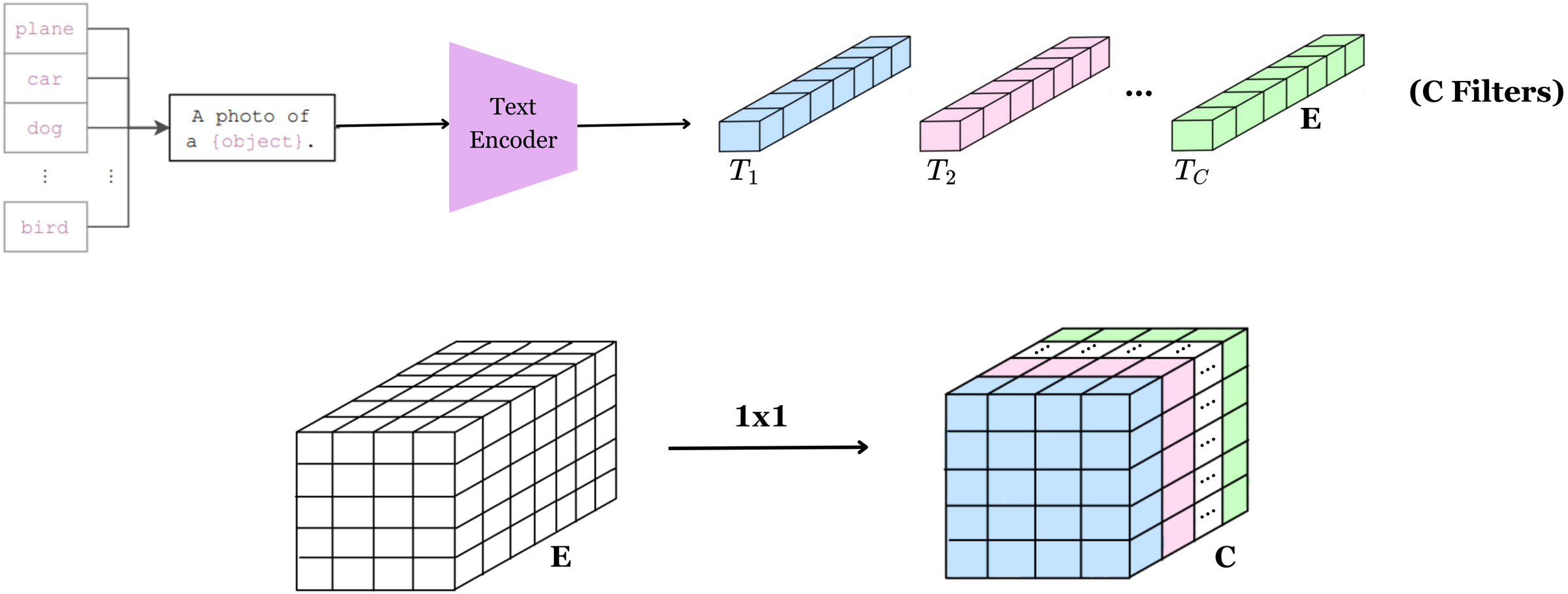
MaskCLIP: CLIP Text Embeddings as Classifier



MaskCLIP: CLIP Text Embeddings as Classifier



MaskCLIP: CLIP Text Embeddings as Classifier



MaskCLIP Refinement Techniques

- KS: Key smoothing

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

- PD: Prompt denoising

MaskCLIP+: Use Segmentation Architectures

Image encoder of CLIP ⟶ Segmentation Architectures

How??

 Backbone + Segmentation Head 

MaskCLIP+: Use Segmentation Architectures

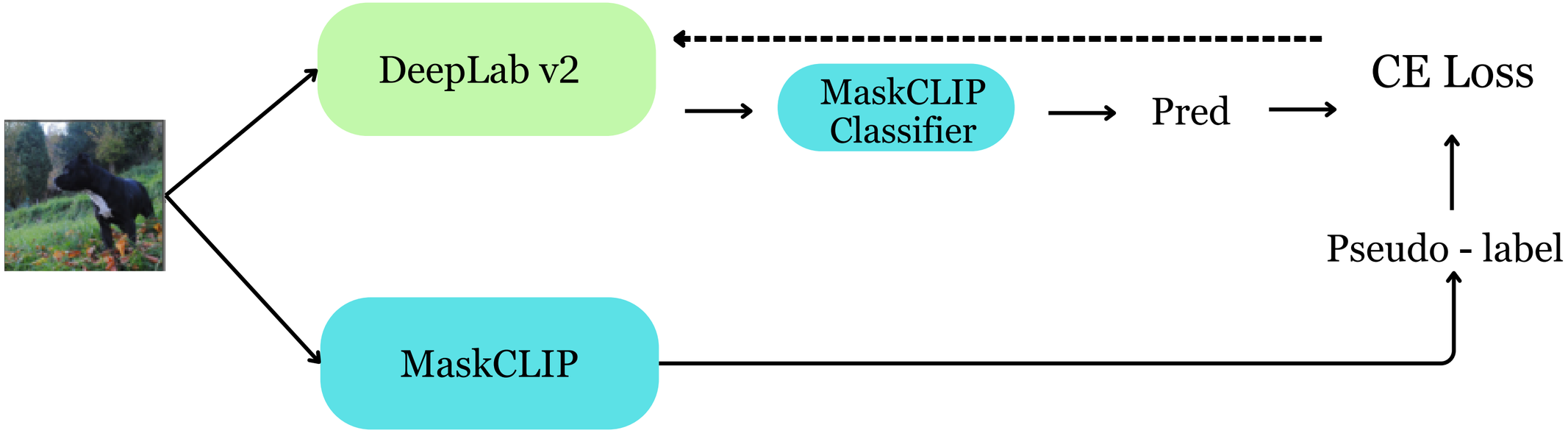
Image encoder of CLIP ⟶ Segmentation Architectures

How??

⇨ Backbone + Segmentation Head ✗

⇨ Use MaskCLIP predictions as a **train-time pseudo GT labels**

MaskCLIP+: Pseudo-labels and Self-trainning

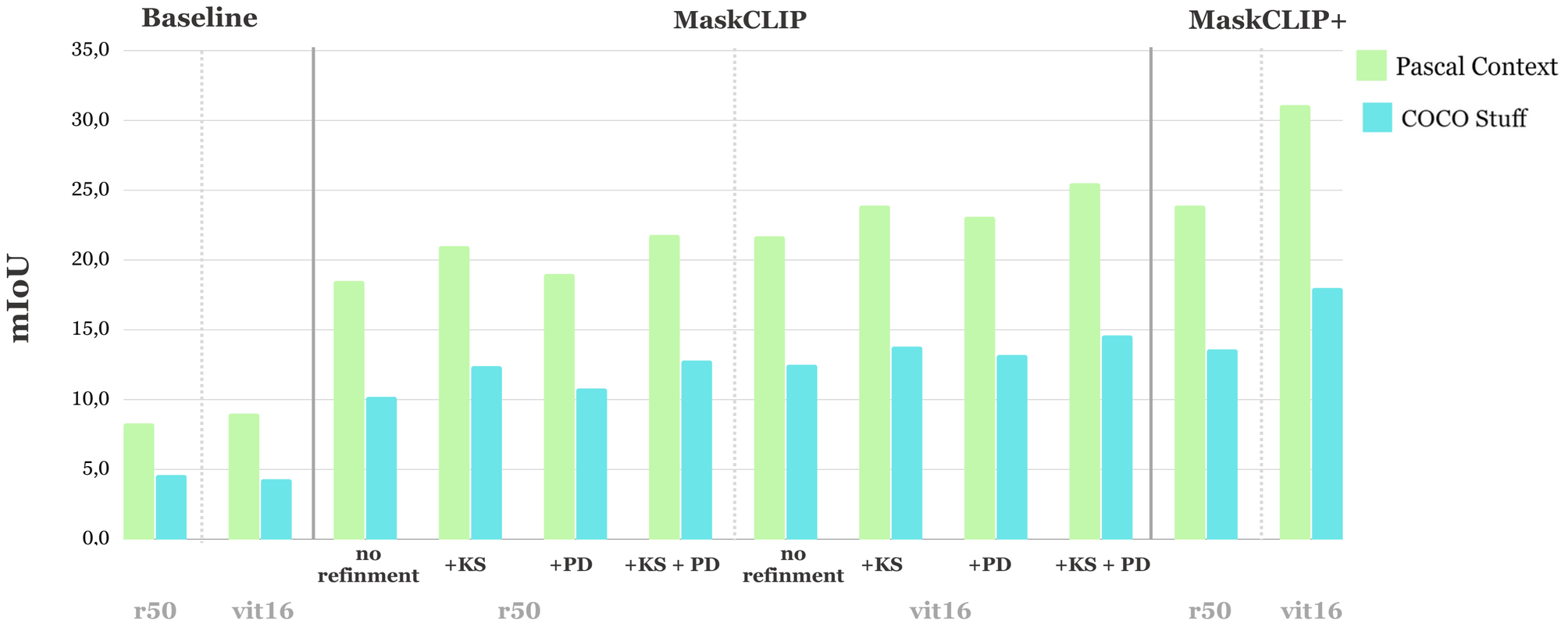


Afer 1/10 train schedule ⇨ Self-Train

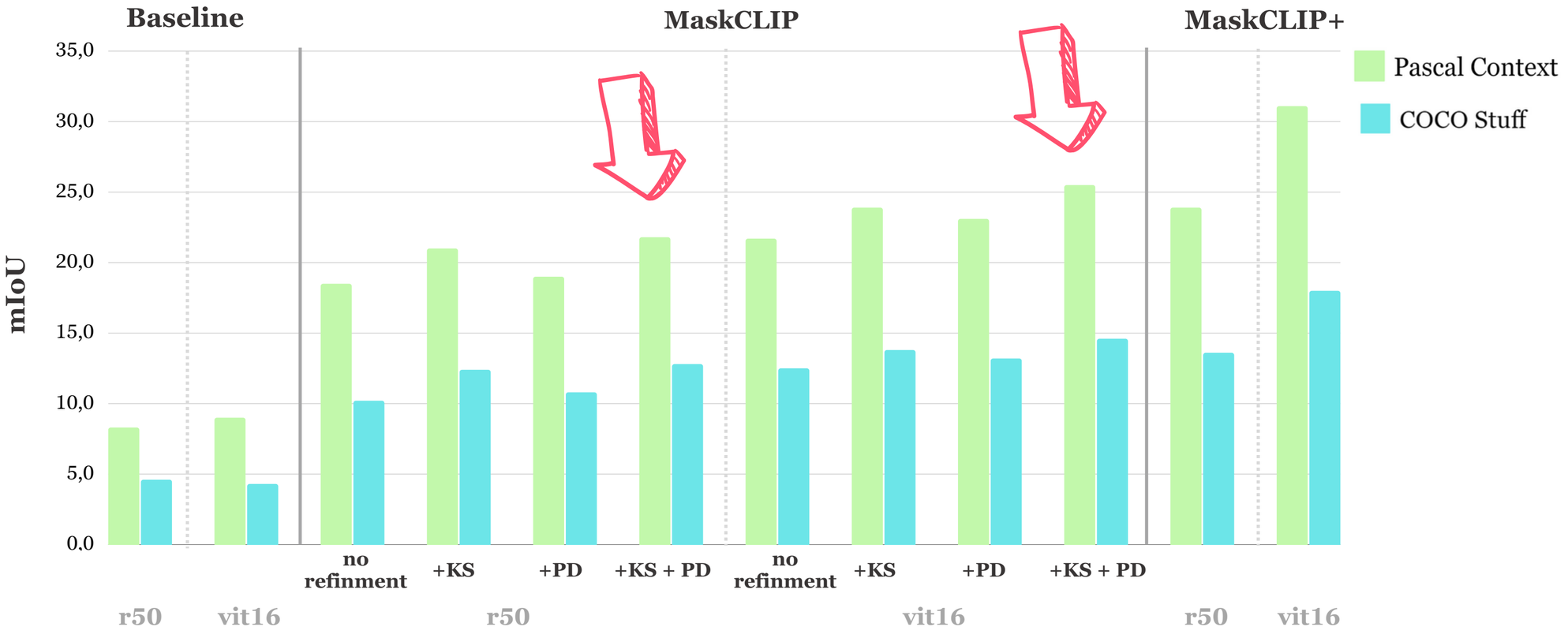
EXPERIMENTS AND RESULTS

Annotation Free Segmentation

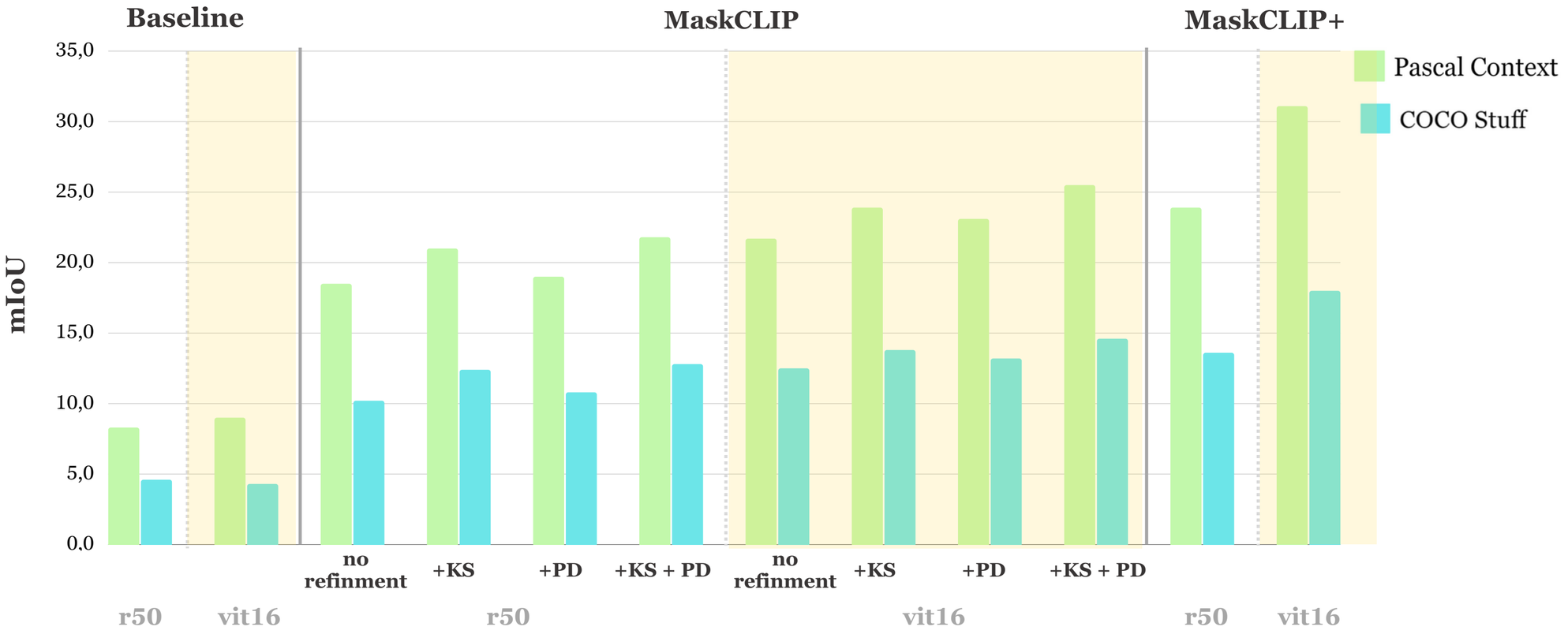
Baseline << MaskCLIP << MaskCLIP+



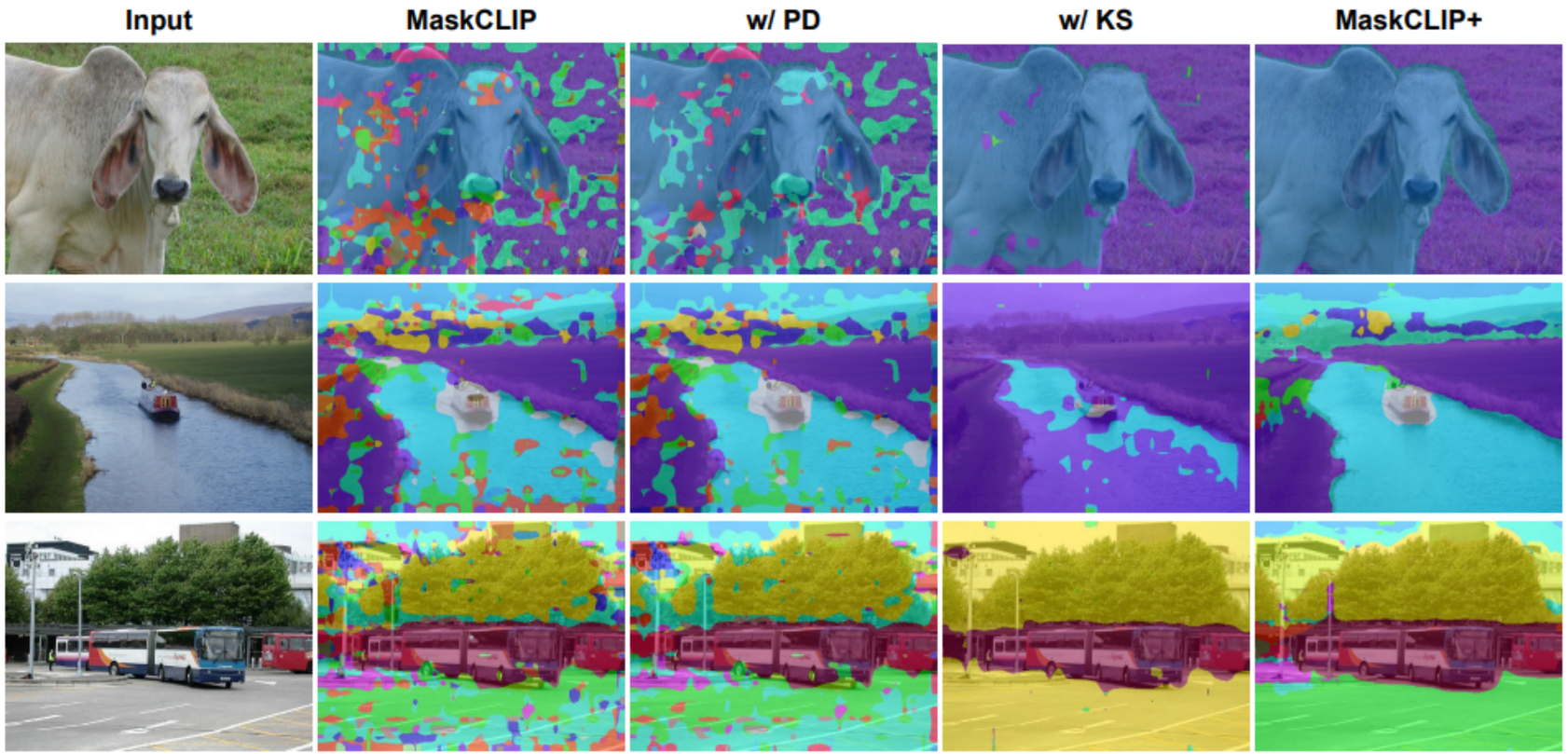
KS and PD are effective and orthogonal



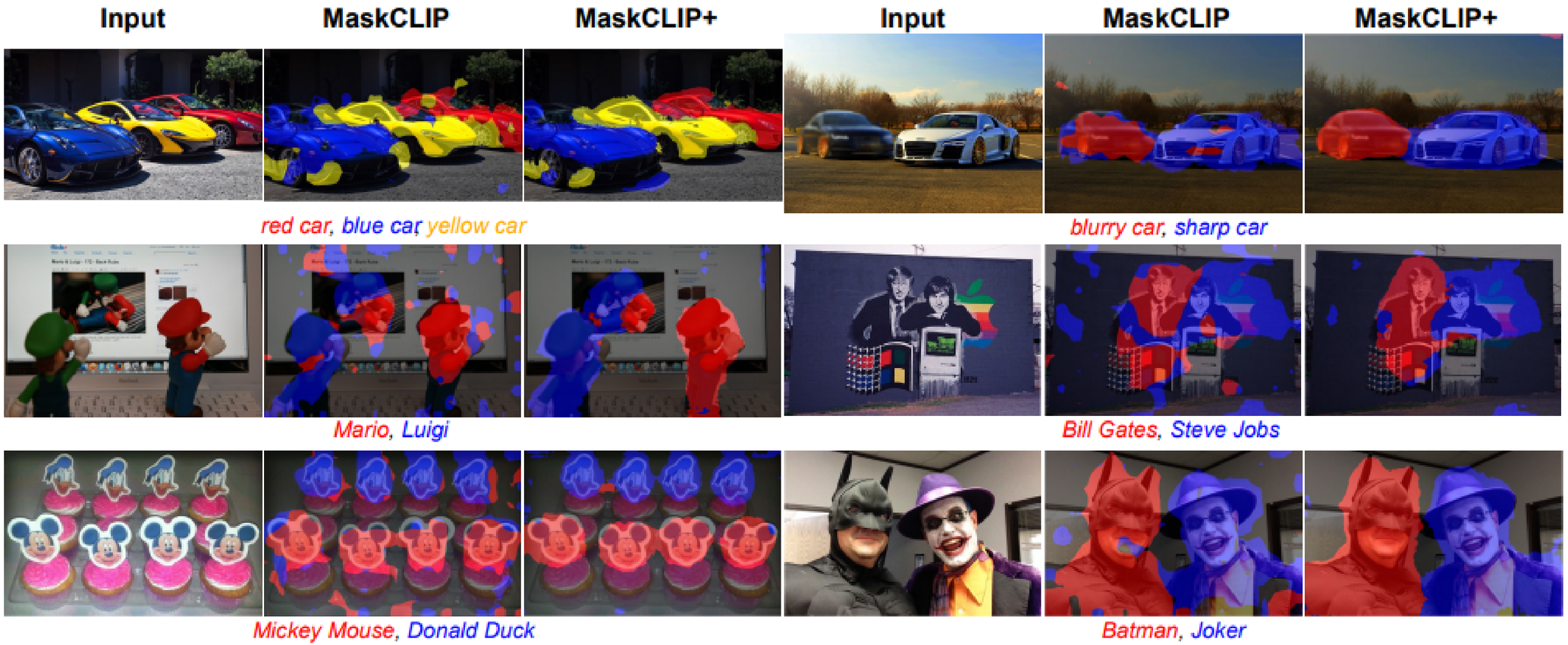
ViT models surpass ResNet models



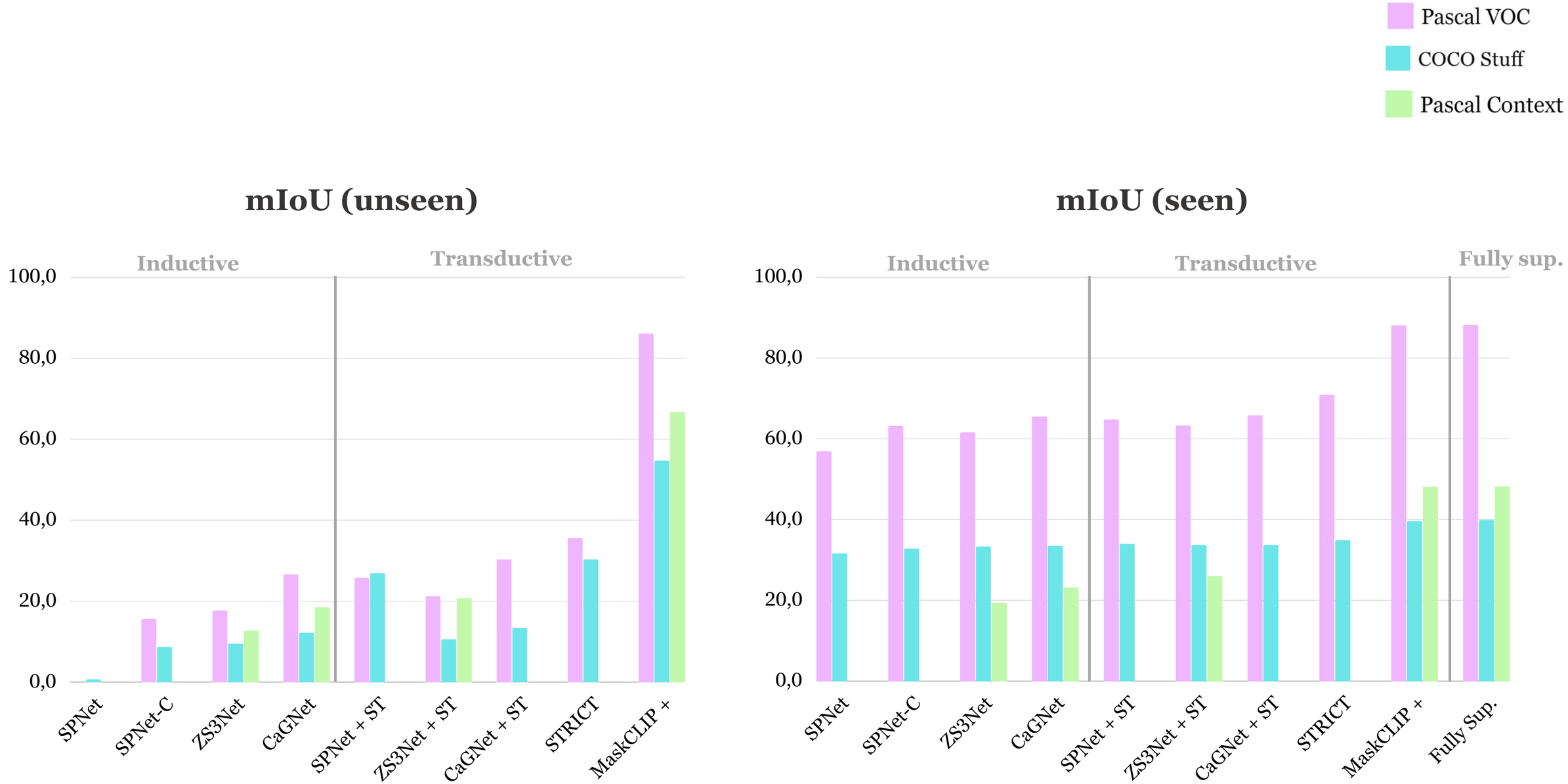
MaskCLIP+ achieves the best performance



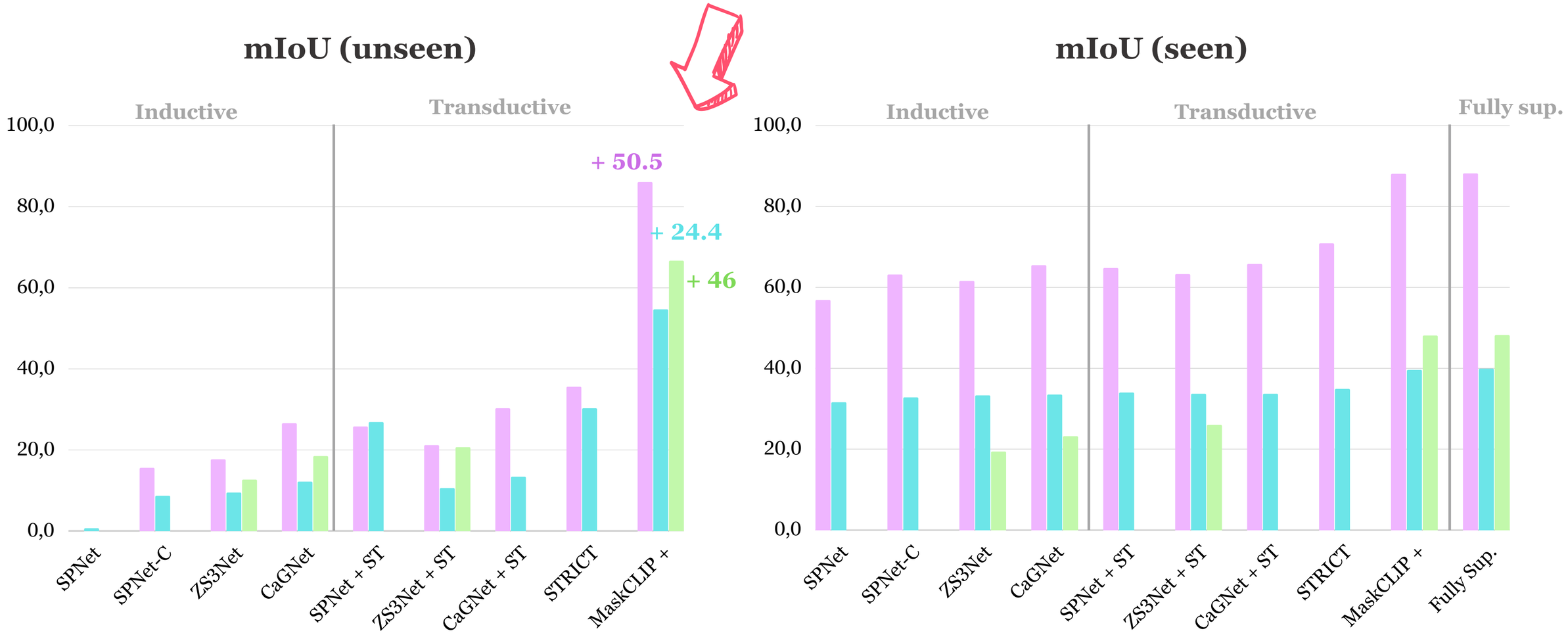
Open-vocabulary Ability



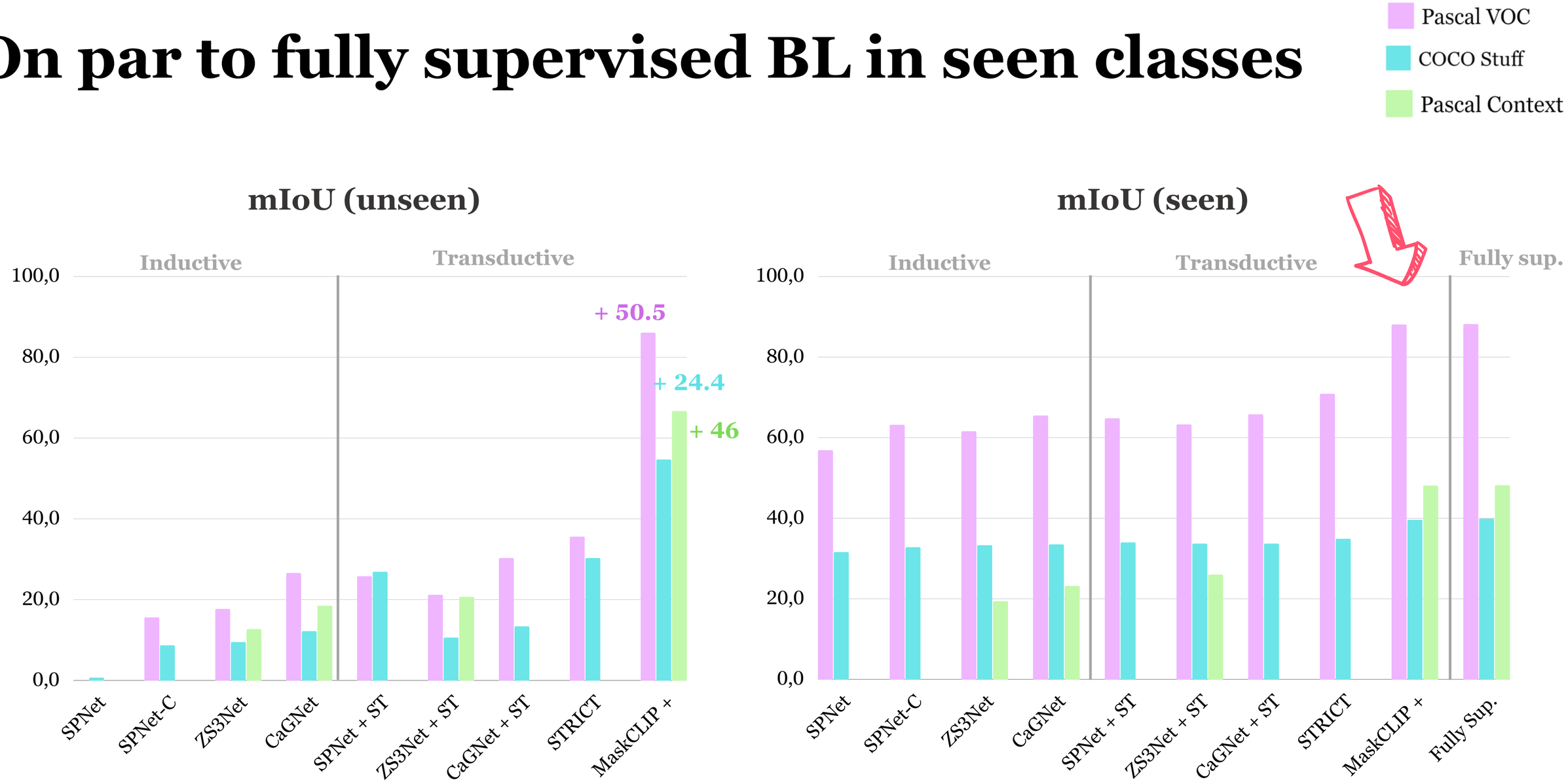
ZeroShot Segmentation


Pascal VOC
COCO Stuff
Pascal Context
mIoU (unseen)
Inductive
Transductive
100,0
80,0
60,0
40,0
20,0
0,0
SPNet
SPNet-C
ZS3Net
CaGNet
SPNet + ST
ZS3Net + ST
CaGNet + ST
STRICT
MaskCLIP +
mIoU (seen)
Inductive
Transductive
Fully sup.
100,0
80,0
60,0
40,0
20,0
0,0
SPNet
SPNet-C
ZS3Net
CaGNet
SPNet + ST
ZS3Net + ST
CaGNet + ST
STRICT
MaskCLIP +
Fully Sup.

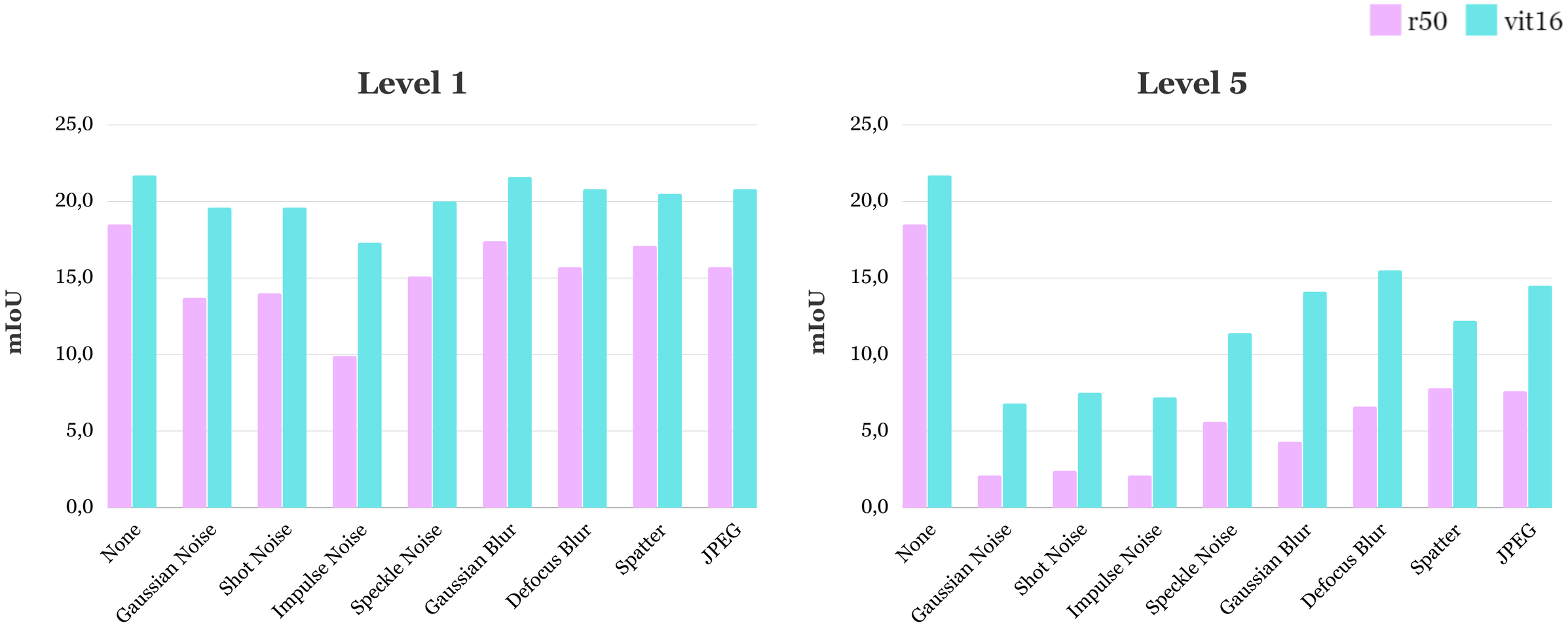
SOTA in unseen classes by large margins



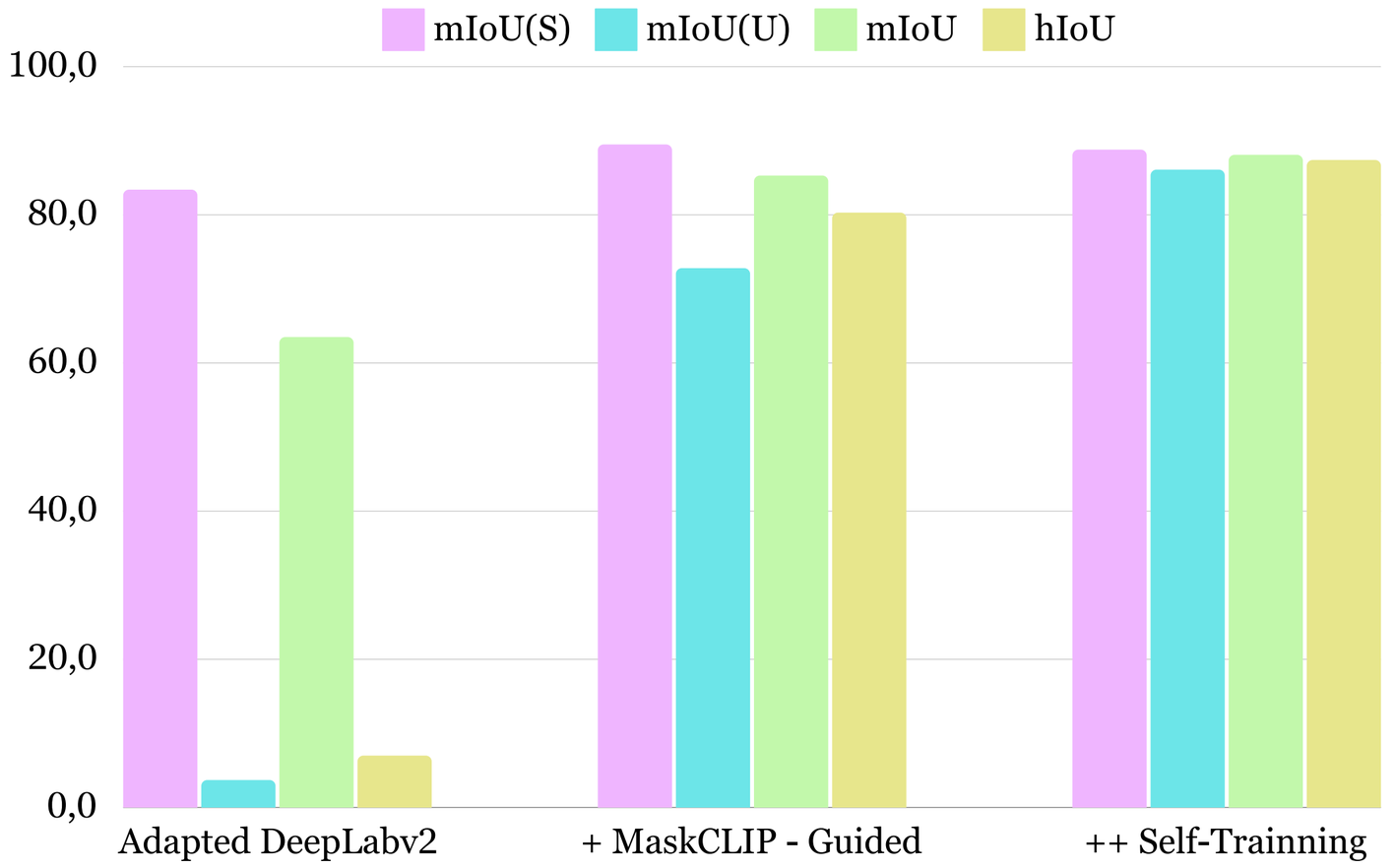
On par to fully supervised BL in seen classes



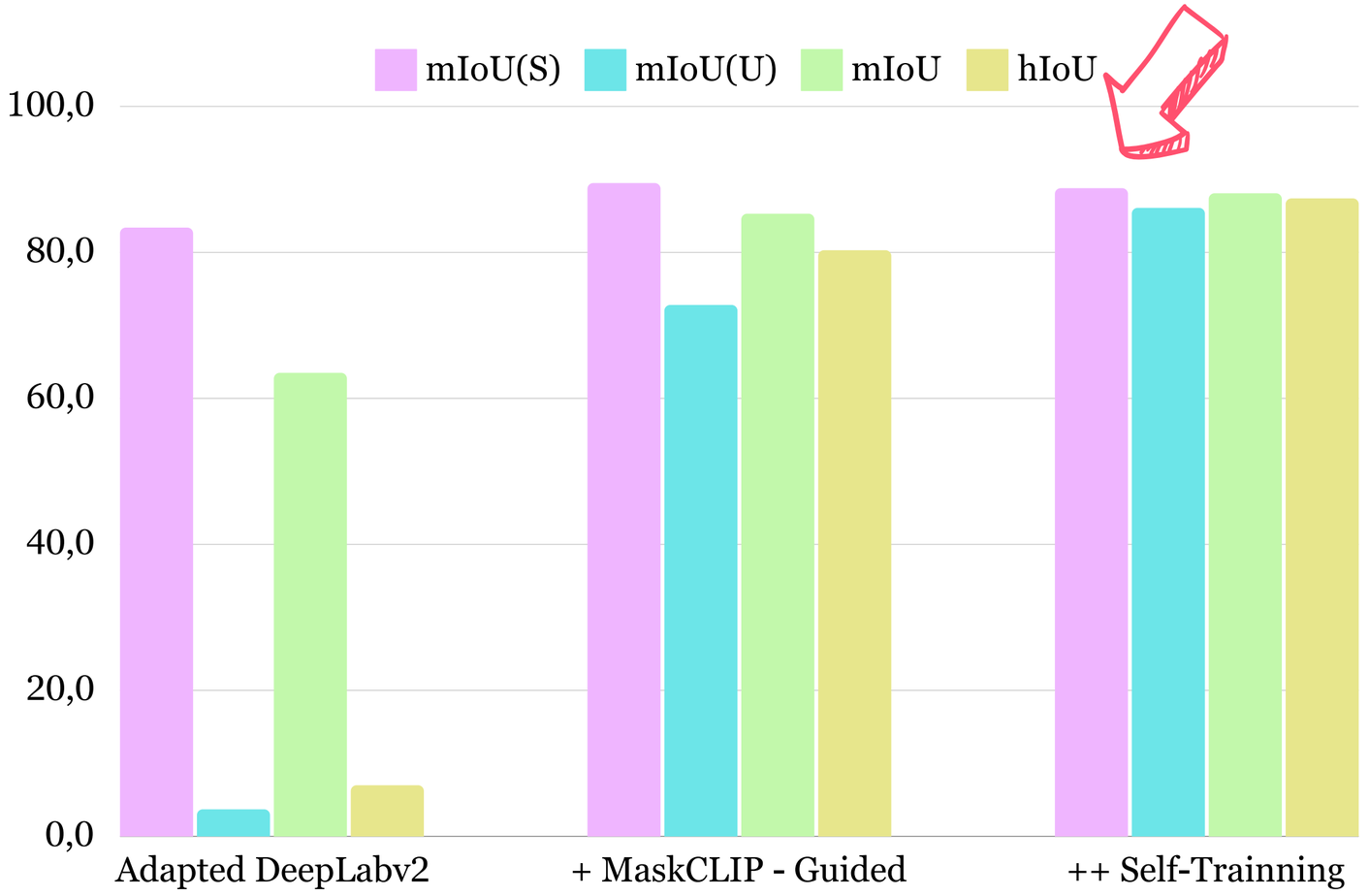
ViT16 models are more robust than ResNet50



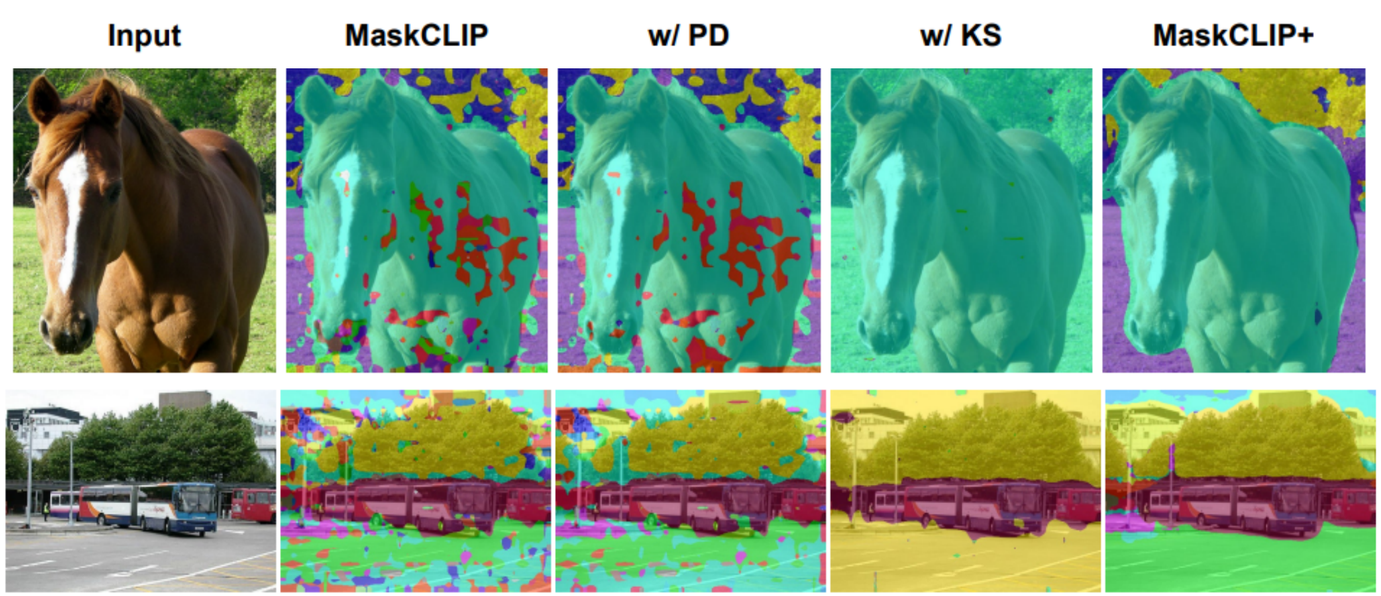
MaskCLIP-guiding and Self-trainning are Effective



Slight degradation on seen classes when Self-training



Failure Cases



MaskCLIP Open Vocabulary

CONCLUSIONS

Architectures Summary

- **MaskCLIP**

  Image encoder: V embedings + 1x1

  Classifier - CLIP Text Embeddings

  KS + PD

- **MaskCLIP+**

  Segmentation model

  MaskCLIP pseduo-labels

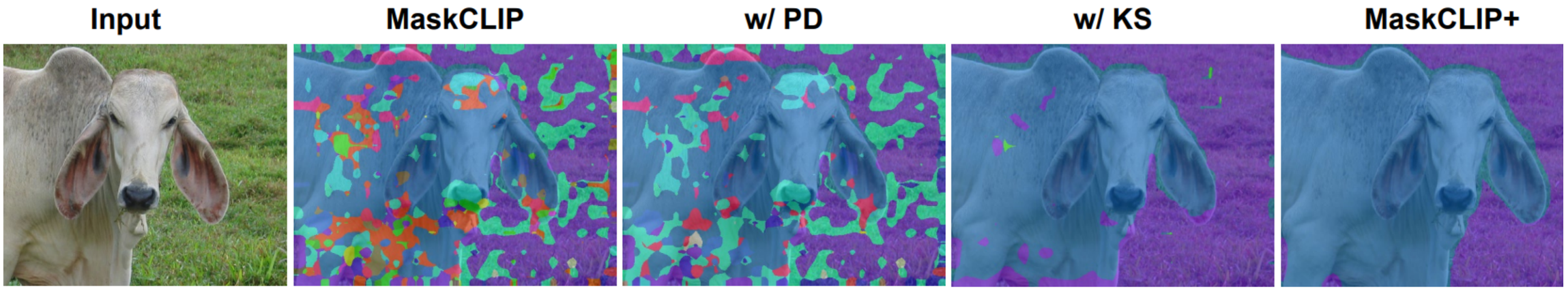
  Self-Train

Towards Cost-Effective Semantic Segmentation

- Expensive Annotations ⇨ Text-image Pairs
  ⇨ MaskCLIP: Free Segmentation Annotator
- Open Vocabulary

Extract Free Dense Labels from CLIP

Authors: Chong Zhou, Chen Change Loy , and Bo Dai



Sonia Castro Paniello
20.02.24