模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

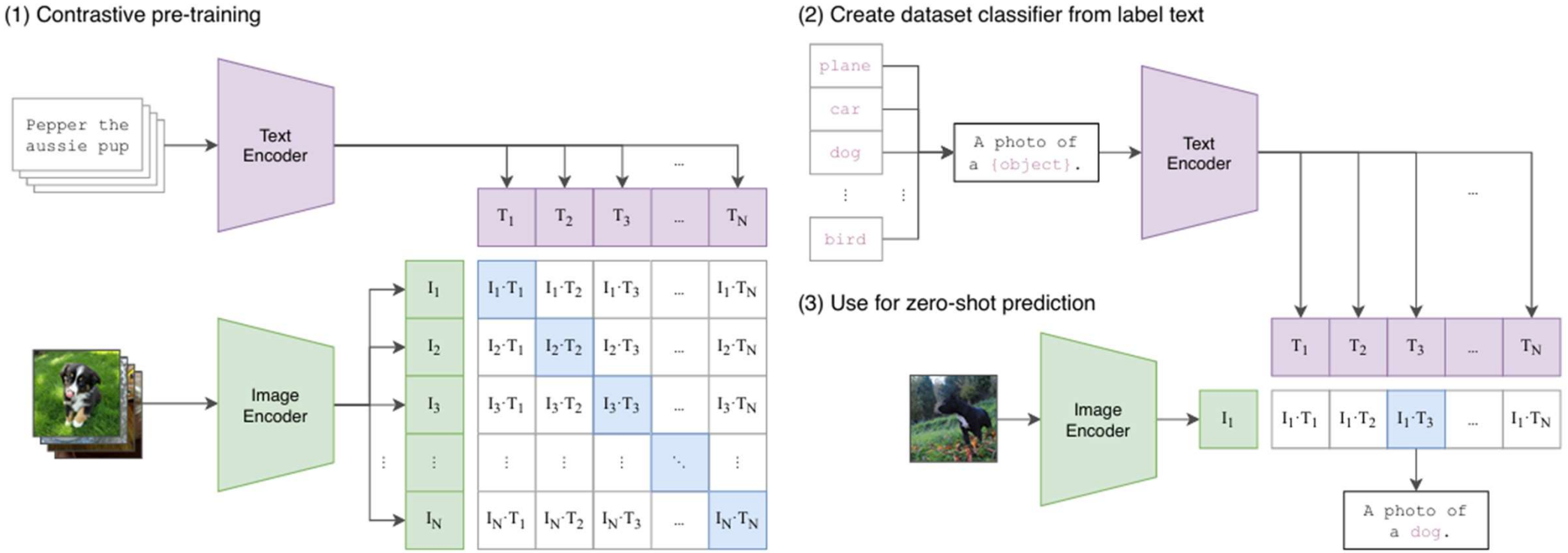
ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

TagCLIP: A Local-to-Global Framework to Enhance Open-Vocabulary Multi-Label Classification of CLIP without Training

Yuqi Lin[1,3], Minghao Chen[2*], Kaipeng Zhang[3*], Hengjia Li[1], Mingming Li[1],
Zheng Yang[4], Dongqin Lv[6], Binbin Lin[5], Haifeng Liu[1], Deng Cai[1,4]

AAAI 2024

CLIP


(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
A photo of a dog.

Motivation

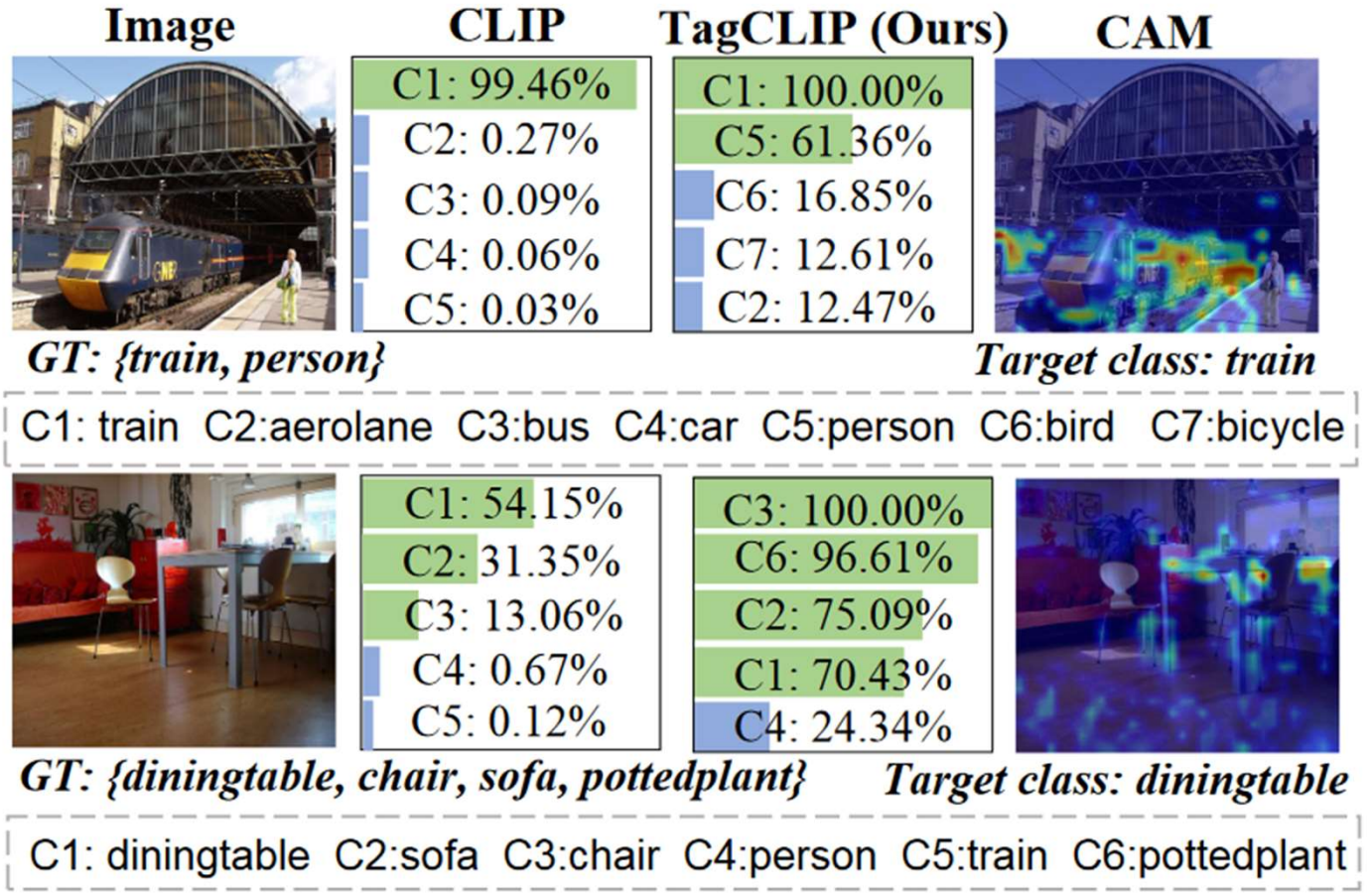


Image
CLIP
TagCLIP (Ours)
CAM
C1: 99.46%
C2: 0.27%
C3: 0.09%
C4: 0.06%
C5: 0.03%
C1: 100.00%
C5: 61.36%
C6: 16.85%
C7: 12.61%
C2: 12.47%
GT: {train, person}
Target class: train
C1: train C2:aerolane C3:bus C4:car C5:person C6:bird C7:bicycle
C1: 54.15%
C2: 31.35%
C3: 13.06%
C4: 0.67%
C5: 0.12%
C3: 100.00%
C6: 96.61%
C2: 75.09%
C1: 70.43%
C4: 24.34%
GT: {diningtable, chair, sofa, pottedplant}
Target class: diningtable
C1: diningtable C2:sofa C3:chair C4:person C5:train C6:pottedplant

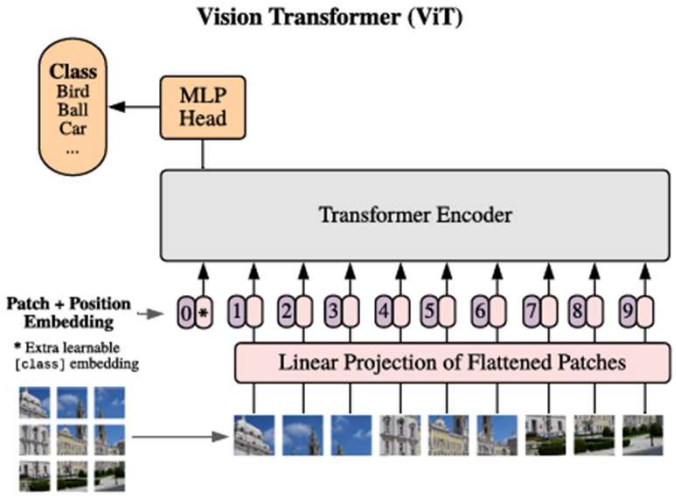
Motivation

The forward propagation of the last transformer layer

$$\begin{aligned}\hat{X}^L &= X^{L-1} + \mathbf{a}^L, \\ &= X^{L-1} + A^L \left(X^{L-1} W_V^L\right), \\ A^L &= \sigma\left(\frac{(X^{L-1}W_Q^L)(X^{L-1}W_V^L)^T}{\sqrt{d}} + M^L\right), \\ X^L &= \hat{X}^L + \mathrm{MLP}(\hat{X}^L),\end{aligned}$$

X^L consists of the [cls] token and remaining tokens (denoted as dense tokens)

$$X^L = [x_{cls}^L, x^L].$$



Let the dense token outputted by the penultimate layer pass the rest layer without self-attention

$$\begin{aligned}\hat{x}_{dense} &= x^{L-1} + \mathbf{c}^L, \\ &= x^{L-1} + x^{L-1} W_V^L, \\ x_{dense} &= \hat{x}_{dense} + \mathrm{MLP}(\hat{x}_{dense}).\end{aligned}$$

Motivation

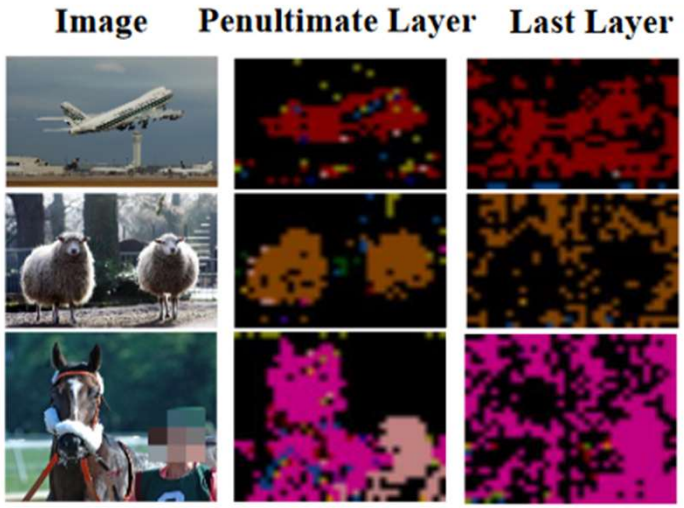
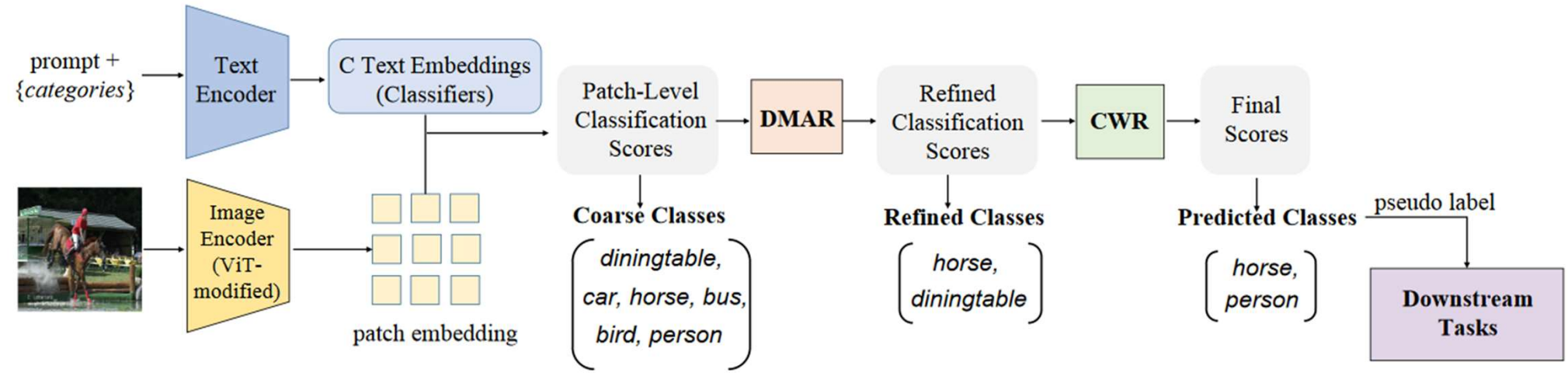


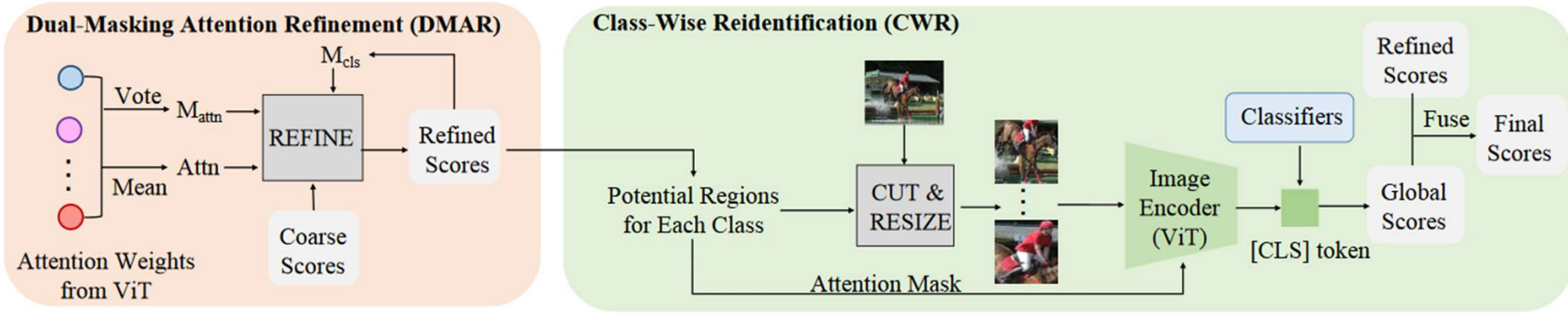
Figure 3: Qualitative results of the patch-level classification upon x_{dense} and x^L outputted by the last two layers of CLIP-ViT respectively. The last self-attention operation breaks spatial information in ViT.

Last Self-Attention	mAP	mIoU
✓	82.7	16.2
✗	85.4	41.6

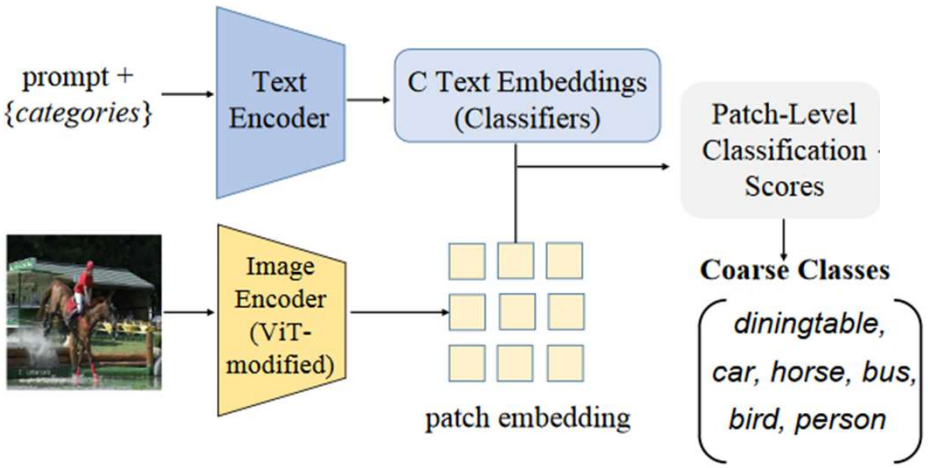
Table 1: Quantitative results for the effect of last self-attention operation in terms of classification (mAP) and segmentation (mIoU) on PASCAL VOC 2012 validation set.

Method


prompt + {categories}
Text Encoder
C Text Embeddings (Classifiers)
Image Encoder (ViT-modified)
patch embedding
Patch-Level Classification Scores
DMAR
Refined Classification Scores
CWR
Final Scores
Coarse Classes
diningtable, car, horse, bus, bird, person
Refined Classes
horse, diningtable
Predicted Classes
horse, person
pseudo label
Downstream Tasks


Dual-Masking Attention Refinement (DMAR)
Vote
Mean
Attn
REFINE
Refined Scores
Coarse Scores
Attention Weights from ViT
Class-Wise Reidentification (CWR)
Potential Regions for Each Class
CUT & RESIZE
Image Encoder (ViT)
Classifiers
[CLS] token
Refined Scores
Fuse
Final Scores
Global Scores
Attention Mask

Coarse Classification

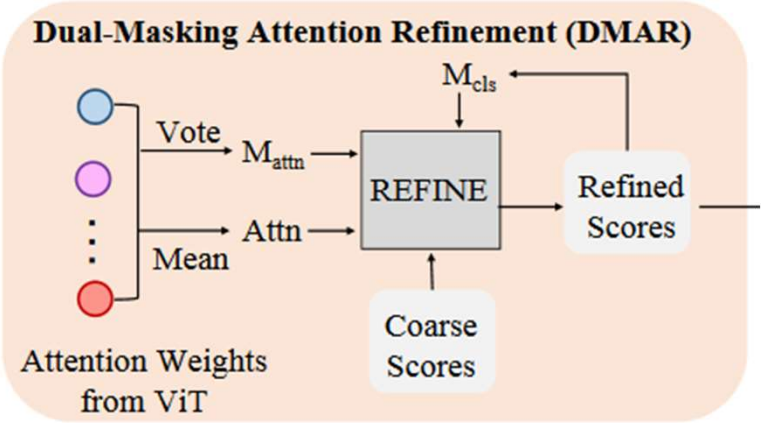


The output feature map based on the penultimate layer $x_{dense} \in R^{N \times D}$ is leveraged. The output of the text encoder is denoted as $T \in R^{D \times C}$, which acts as the classifier based on the text inputs. N, D, C represent token length, token dimension and class number, respectively.

$$s_i = Linear(x_{dense,i}) * T,$$

$$P_{coarse}(i,c) = \frac{\exp(s_i^c)}{\sum_{k=1}^{C} \exp(s_i^k)}.$$

Dual-Masking Attention Refinement (DMAR)



$$P_{refined} = \frac{1}{|\psi|} \sum_{l \in \psi} A_l * P_{coarse},$$

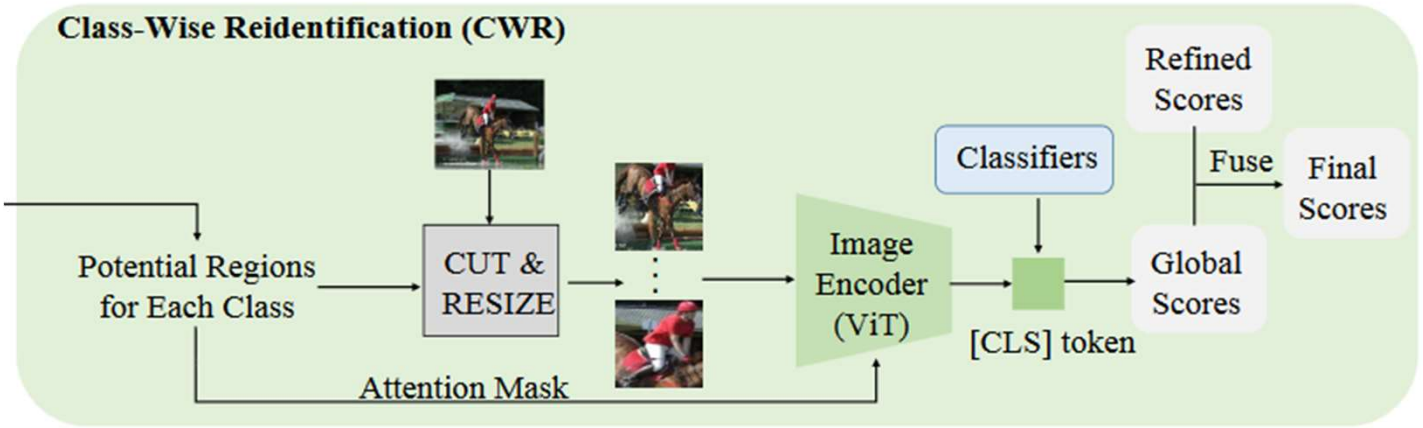
$P_{coarse} \in R^{N \times C}$ denotes the coarse score map, $A_l \in R^{N \times N}$ represents the attention weight in the l-th layer of ViT, ψ represents the index set of used attention layer and $|\psi|$ is its number of elements.

$$M_{attn}(i,j) = 1, if \sum_{l=1}^{L} \mathbb{I}_{(A(i,j,l) > \bar{A}_l)}(A) > K,$$

$$\hat{P}_{refined} = \frac{1}{|\psi|} \sum_{l \in \psi} M_{attn} \odot A_l * P_{coarse},$$

$$P_{refined}(c) = \frac{1}{|\psi|} \sum_{l \in \psi} M_{attn} \odot A_l \odot M_{cls}(c) * P_{coarse}(c).$$

Class Scores

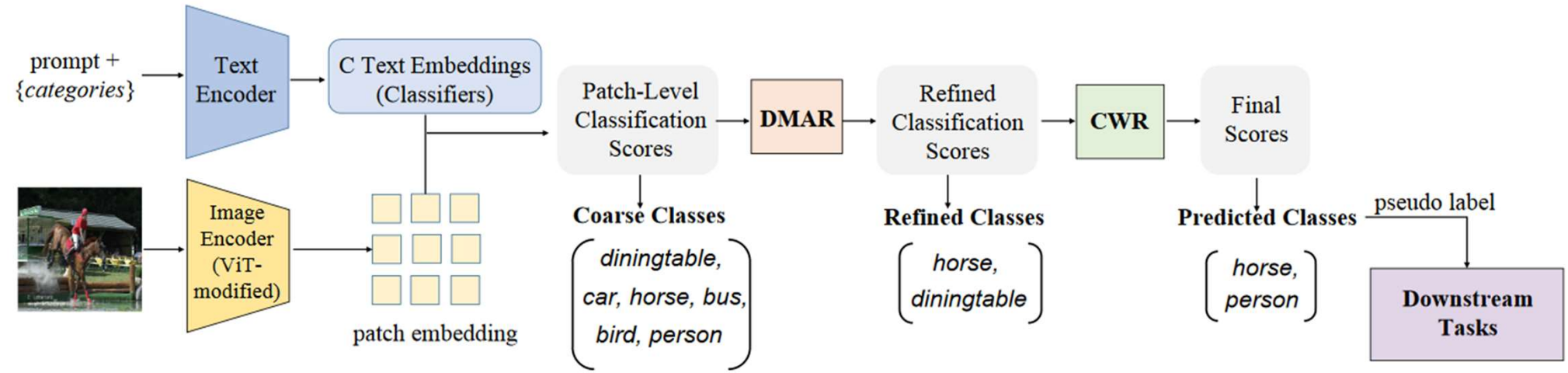


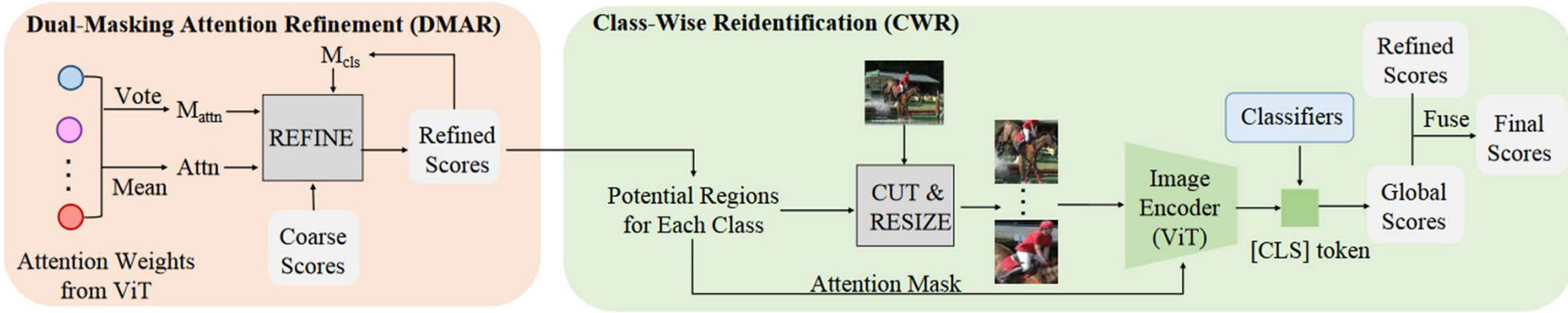
$$P_{local}(c) = \max_i(P_{coarse}(i, c)),$$

For each class, we pick out the highly responsive patches from $P_{refined}$ and form the class-related region (class-wise mask). We crop the image by the bounding box of the region and resize it to a specific size, e.g., 224×224. The class-wise mask serves as the attention mask in ViT to exclude patches that do not belong to the class. We input the class-wise image into original CLIP and use [cls] token for classification.

$$P_{final} = \lambda P_{local} + (1 - \lambda) P_{global},$$

Method


prompt + {categories}
Text Encoder
C Text Embeddings (Classifiers)
Image Encoder (ViT-modified)
patch embedding
Patch-Level Classification Scores
DMAR
Refined Classification Scores
CWR
Final Scores
Coarse Classes
diningtable, car, horse, bus, bird, person
Refined Classes
horse, diningtable
Predicted Classes
horse, person
pseudo label
Downstream Tasks


Dual-Masking Attention Refinement (DMAR)
Vote
Mean
Attn
REFINE
Refined Scores
Coarse Scores
Attention Weights from ViT
Class-Wise Reidentification (CWR)
Potential Regions for Each Class
CUT & RESIZE
Image Encoder (ViT)
Classifiers
[CLS] token
Refined Scores
Fuse
Final Scores
Global Scores
Attention Mask

Experiment

Method	Extra Training Data	VOC	COCO
Supervised specialist:			
SARB	10% Data	83.5	75.5
DualCoOp	10% Data	90.3	78.7
TAI-DPT	10% Data	93.3	81.5
Open-vocabulary generalist:			
TAI-DPT	COCO captions	88.3	65.1
CLIP†	None	79.5	54.2
CLIP	None	85.8	63.3
DPT†	None	83.4	59.6
DPT	None	86.2	64.3
CLIPSurgery	None	85.4	61.2
TagCLIP(Ours)	None	**92.8**	**68.8**

Table 2: Experimental results of multi-label classification. † represents not using softmax on classification scores.

Method	VOC	COCO	COCO-Stuff
Vanilla USS methods			
IIC	9.8	-	6.7
MaskContrast	35.0	3.73	-
TransFGU	37.2	12.7	17.5
MaskDistill	45.8	-	-
PiCIE	-	-	13.8
PiCIE+H	-	-	14.4
CLIP-based methods			
MaskCLIP‡	42.1	20.2	23.9
CLIPSurgery‡	41.5	25.2	29.7
GroupViT	52.3	24.3	-
SegCLIP	52.6	26.5	-
ReCo	34.2	17.1	26.3
NamedMask	59.2	27.7	-
CLS-SEG (Ours)	**64.8**	**34.0**	**30.1**
CLS-SEG*(Ours)	**68.7**	**35.3**	**31.0**

Table 3: Results of annotation-free semantic segmentation. The vanilla USS results are based on K-means clustering. ‡ represents we re-implement it with the same experimental setting as ours. * means using denseCRF to postprocess.

Experiment

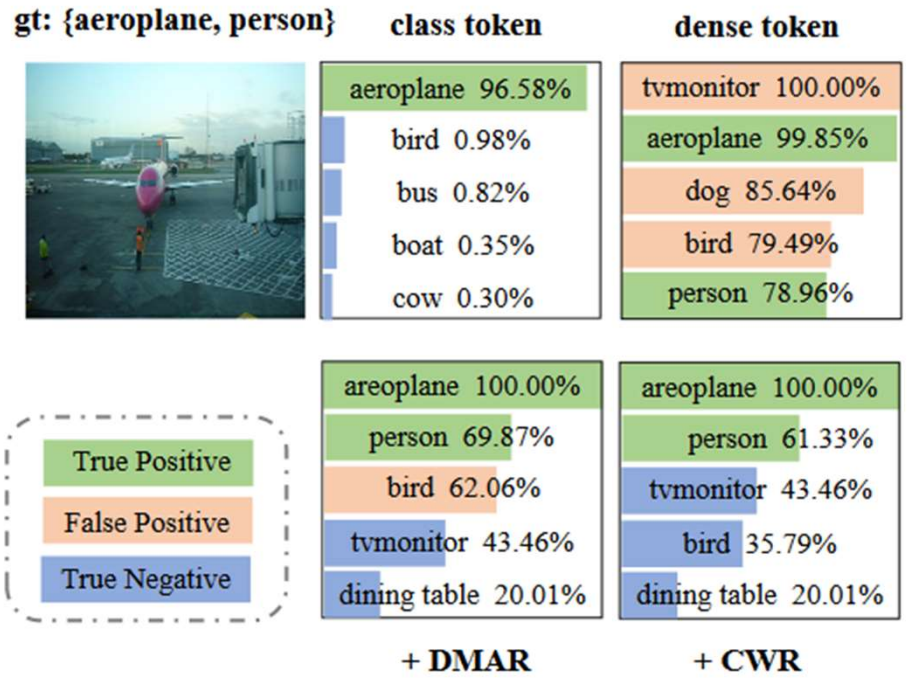
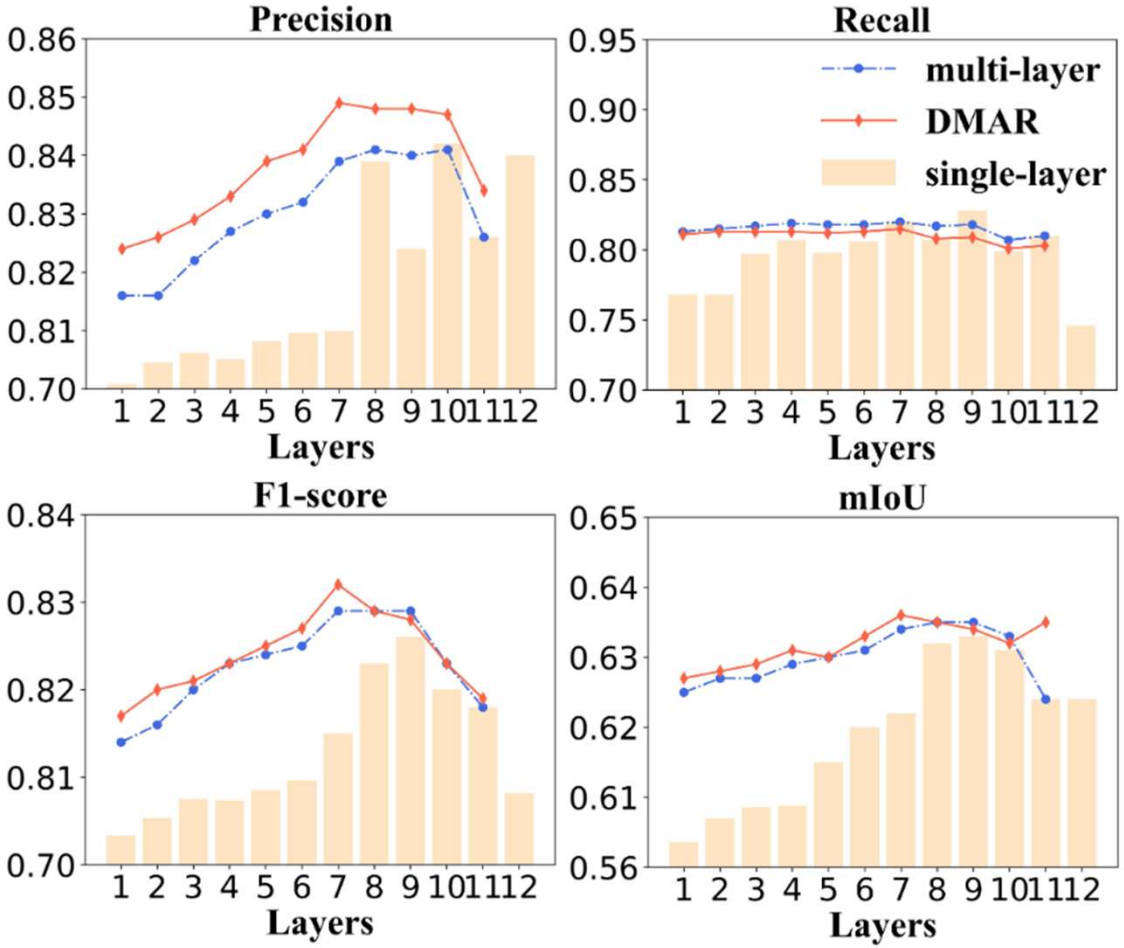


Figure 5: The classification results of different strategies. We use 0.5 as the threshold by default.

Coarse Score	DMAR	CWR	mAP	mIoU
✓			85.4	30.9
✓		✓	88.0	55.2
✓	✓		93.9	63.7
✓	✓	✓	94.1	64.8

Table 4: Results for the effectiveness of DMAR and CWR module in terms of classification and semantic segmentation. The results are evaluated on the VOC 2012 val set.

Experiment



Comparison of single-layer and multi-layer attention refinement in terms of classification and segmentation tasks. For the single-layer setting, each tick i on the x-axis represents merely adopting attention weight in i-th layer. For the multi-layer setting, the i-th x-tick means fusing i-th to 11-th layers attention weights to refine coarse classification scores. We rule out the last attention layer during fusing.

Thanks