南京航空航天大學

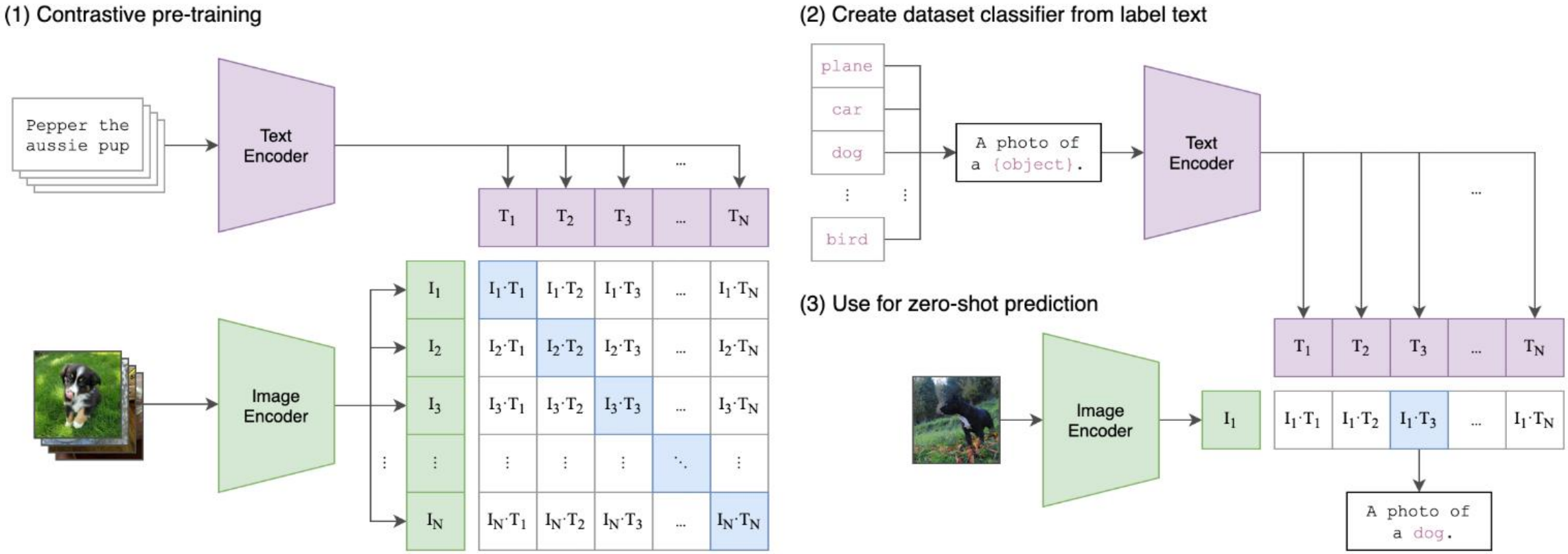
microCLIP: Unsupervised CLIP Adaptation via Coarse-Fine Token Fusion for Fine-Grained Image Classification

Sathira Silva[1] Eman Ali[1,2] Chetan Arora[3] Muhammad Haris Khan[1]

CVPR 2025

Introduction

解决问题：CLIP在细粒度分类任务上表现较差



CLIP 在 zero-shot 分类中表现很强，但在细粒度任务上性能受限

原因：CLIP 的 [CLS] token 是全局特征，忽略了图像中局部的、判别性的细节

Introduction

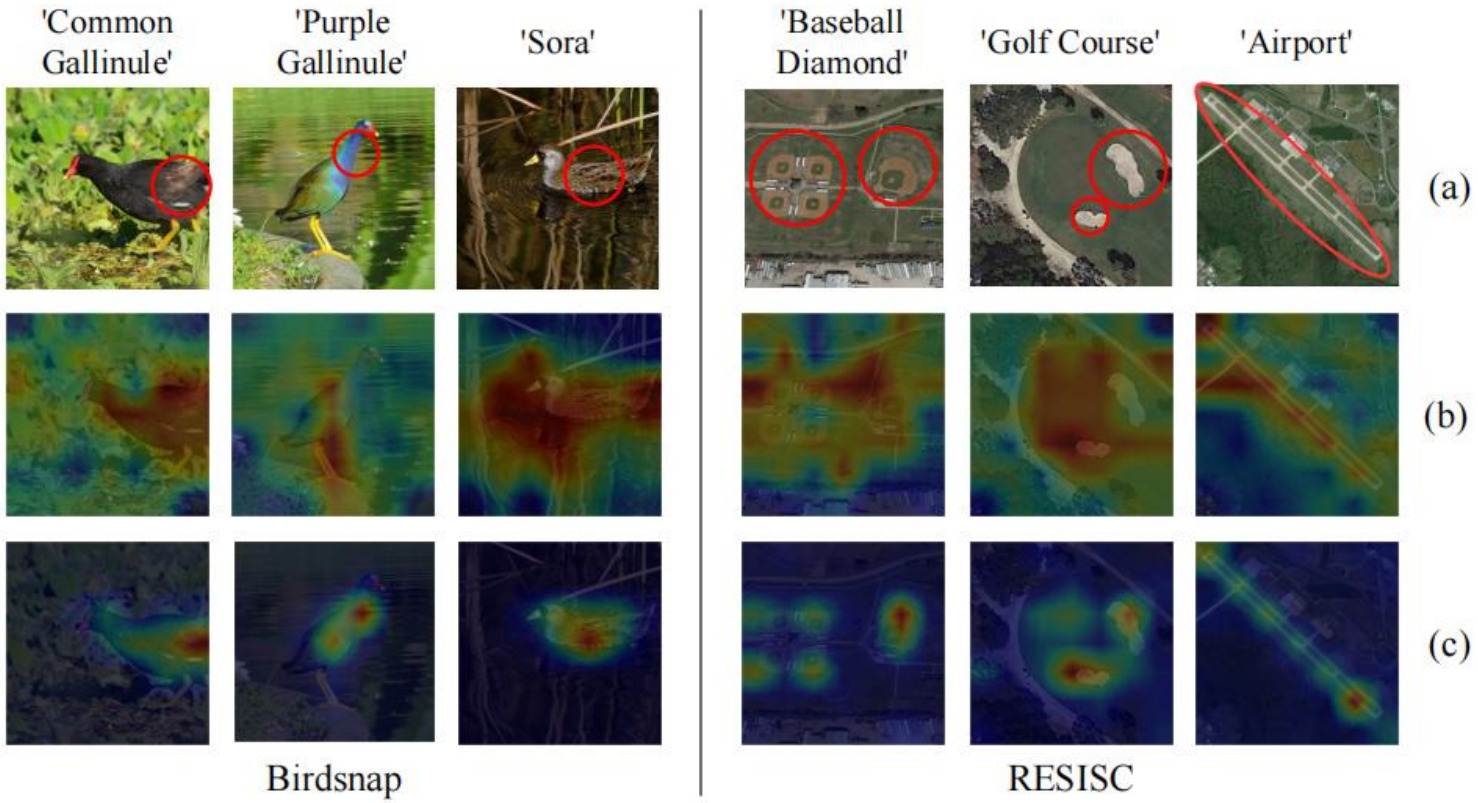


Figure 1. Attention maps on two fine-grained datasets: Birdsnap and RESISC45. Row (a): input images; (b): global attention from DPA [2]; (c): local attention from **microCLIP** (ours). By guiding the [FG] token with SOAP queries, microCLIP focuses on semantically critical regions, yielding sharper, more discriminative attention. Red circles highlight referenced regions in the text.

DPA: Dual Prototypes Alignment for Unsupervised Adaptation of Vision-Language Models

Related Works

南京航空航天大學

无监督自适应 (Unsupervised Adaptation of CLIP)
UPL\POUF\LaFTer无监督微调。ReCLIP投影空间解决视觉-文本错位问题，并同时微调两个编码器，同时投入大量成本进行伪标记的标签传播，成本很高。
DPA从无标签的图片中找出一些图像原型，然后让文本描述和这些“代表图片”对齐,减少错误。但它还是依赖于 CLIP 的 [CLS] token，而这个 token 是全局的，不够精细。

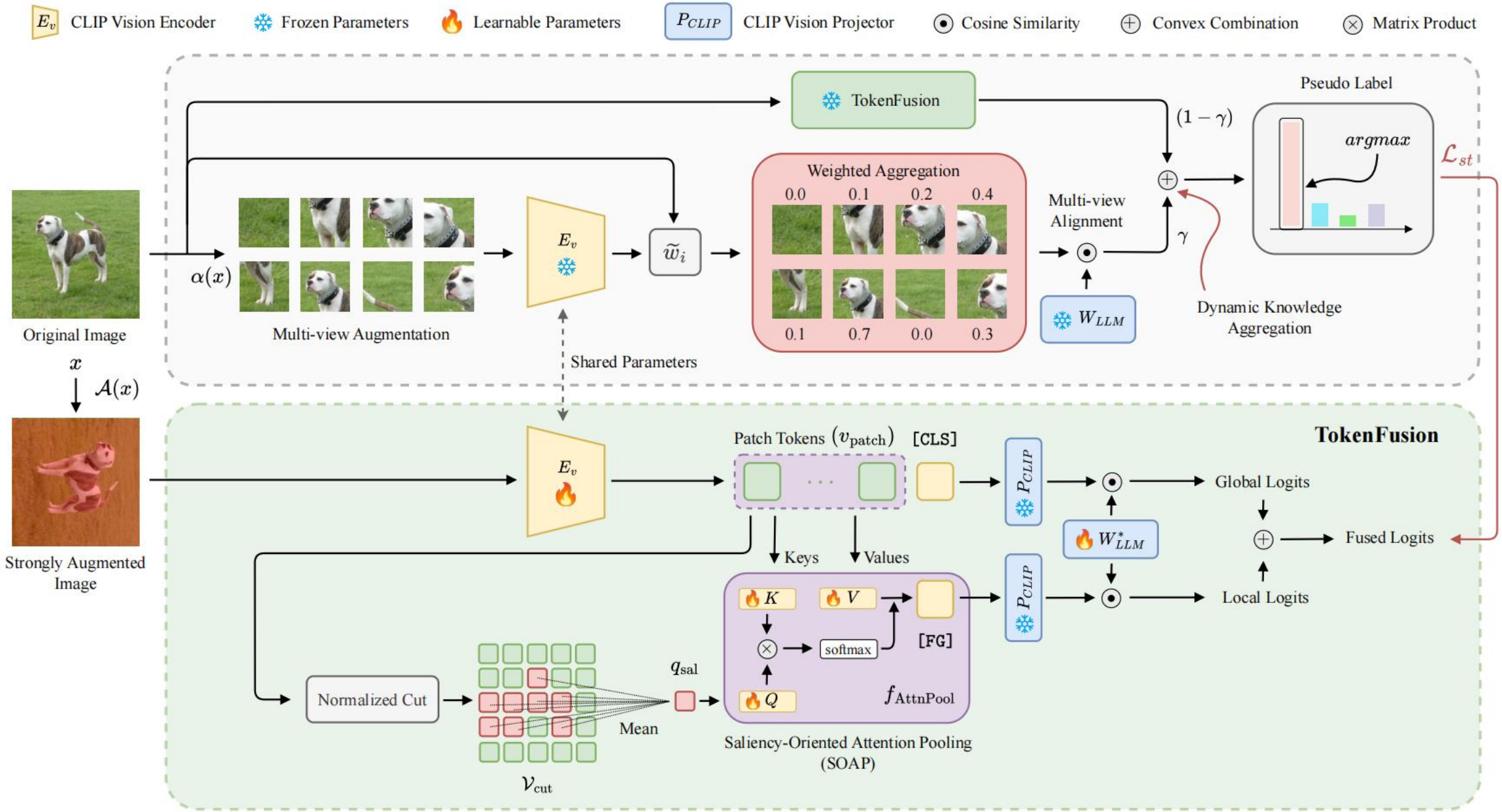
多视角表示 (Multi-view Representations)
用于生成伪标签的增强视图
DINO-MC, VCR：这些方法会切出不同大小的图片块，来丰富 CLIP 的特征。
WCA (这是 microCLIP 借鉴灵感的对象)：它会随机切出非常多（比如60个）的小块，然后把这些小块的预测结果加权平均，得到一个更可靠的判断。计算量巨大。本方法只切很少的几个块（比如8个）

提取显著区域 (Extraction of Salient Regions)
采用Normalized Cut (NCut)。将图片里的 patch tokens分成两组，一组是最重要的，一组是次要的。

细粒度任务中的 Patch Tokens 利用
TagCLIP： CLIP 视觉编码器的倒数第二层的 patch tokens 比最后一层的更优。最后一层的 patch tokens 已经被 [CLS] token “污染”，失去了局部信息，而倒数第二层还保留着清晰的空间细节。

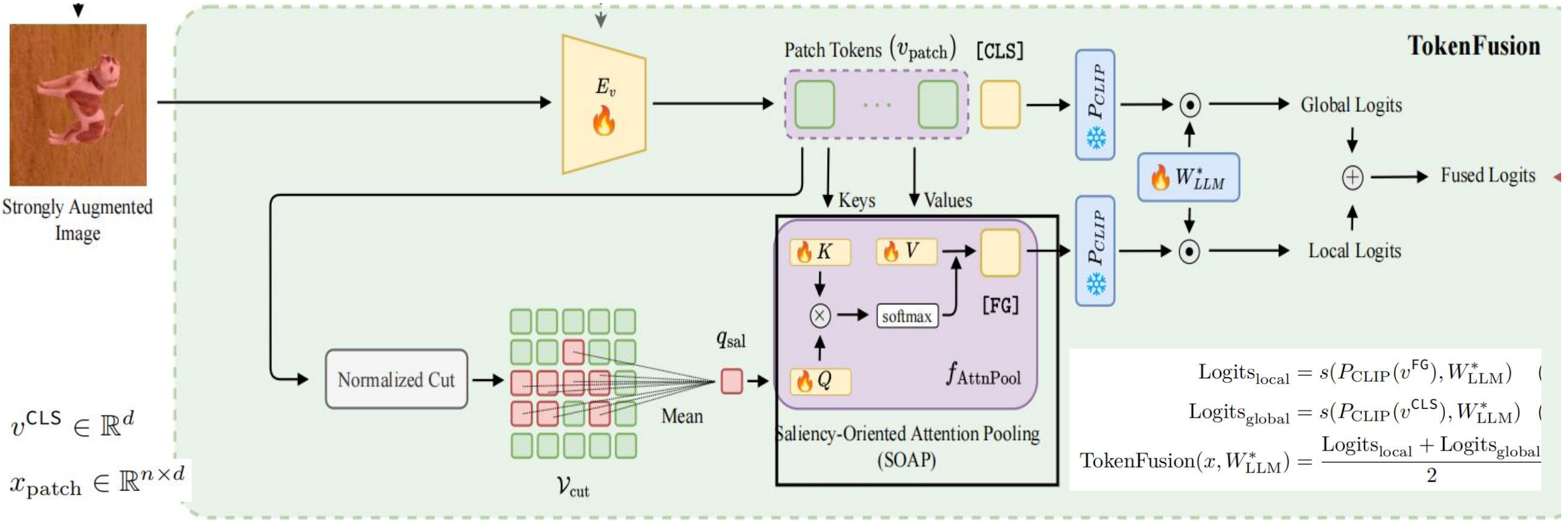
clip_skip = 1	使用倒数第 2 层（Layer N-1）输出；，保留关键词准确度	输出层不同，对 prompt 理解深浅不同，直接影响最终图像生成内容。
clip_skip = 2	使用倒数第 3 层（Layer N-2）输出；最大限度还原 prompt 精度	
clip_skip = 0	或默认 → 使用最后一层（Layer N）输出，风格高度融合，但可能失真	

Method


E_v CLIP Vision Encoder
Frozen Parameters
Learnable Parameters
P_{CLIP} CLIP Vision Projector
Cosine Similarity
Convex Combination
Matrix Product
TokenFusion
Pseudo Label
$(1-\gamma)$
γ
$\mathcal{L}_{st}$
argmax
Weighted Aggregation
0.0 0.1 0.2 0.4
0.1 0.7 0.0 0.3
Multi-view Alignment
W_{LLM}
Dynamic Knowledge Aggregation
$\alpha(x)$
$\widetilde{w}_i$
Multi-view Augmentation
Original Image
x
$\mathcal{A}(x)$
Strongly Augmented Image
Shared Parameters
TokenFusion
Patch Tokens (v_{patch})
[CLS]
Keys
Values
W^*_{LLM}
Global Logits
Fused Logits
Local Logits
K
V
Q
softmax
[FG]
f_{AttnPool}
Saliency-Oriented Attention Pooling (SOAP)
Normalized Cut
q_{sal}
Mean
$\mathcal{V}_{\text{cut}}$

Method TokenFusion

Saliency-Oriented Attention Pooling（SOAP）面向显著性的注意力池化



$v^{\text{CLS}} \in \mathbb{R}^d$

$x_{\text{patch}} \in \mathbb{R}^{n \times d}$

$$\text{Logits}_{\text{local}} = s(P_{\text{CLIP}}(v^{\text{FG}}), W^*_{\text{LLM}})$$

$$\text{Logits}_{\text{global}} = s(P_{\text{CLIP}}(v^{\text{CLS}}), W^*_{\text{LLM}})$$

$$\text{TokenFusion}(x, W^*_{\text{LLM}}) = \frac{\text{Logits}_{\text{local}} + \text{Logits}_{\text{global}}}{2}$$

$$E_v(x) = [x_{\text{patch}}, v^{\text{CLS}}]$$

$$\tilde{v}_{\text{patch}} = x^{L-1}_{\text{patch}} + x^{L-1}_{\text{patch}} \widetilde{W}^L_V$$

$$v_{\text{patch}} = \tilde{v}_{\text{patch}} + \text{MLP}(\tilde{v}_{\text{patch}})$$

$$\mathcal{V}_{\text{cut}} = \text{NCut}(v_{\text{patch}})$$

$$q_{\text{sal}} = \frac{1}{|\mathcal{V}_{\text{cut}}|} \sum_{\forall v \in \mathcal{V}_{\text{cut}}}$$

$$v^{\text{FG}} = f_{\text{AttnPool}}(q_{\text{sal}}, v_{\text{patch}})$$

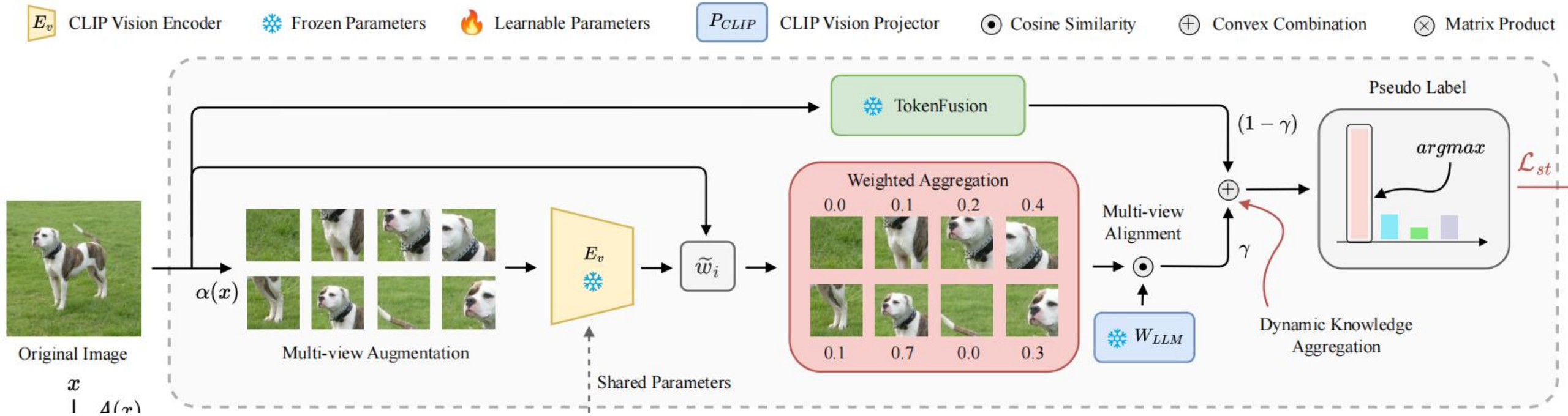
$$= \text{softmax}\left(\frac{q_{\text{sal}} W_{\text{Q}} (v_{\text{patch}} W_{\text{K}})^\top}{\sqrt{d}}\right) v_{\text{patch}} W_{\text{V}}$$

Method

通过动态知识聚合迭代优化伪标记

Iteratively Improving Pseudo-Labels with Dynamic Knowledge Aggregation

南京航空航天大學



$$\alpha(x) = \{x_i | x_i = \phi(x, \lambda_i \min(H, W)) \mid i = 1 \dots N\}$$

$$w_i = \frac{\exp\left(s(f(x), f(x_i))\right)}{\sum_{l=1}^{N} \exp\left(s(f(x), f(x_l))\right)}$$

$$f^{\text{agg}}(x) = \sum_{i=1}^{N} w_i \cdot f(x_i|\alpha)$$

$$\text{Pseudo-logits}_{\text{CLIP}} = s(f^{\text{agg}}(x), W_{\text{LLM}}|\alpha)$$

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} \{\gamma \cdot \text{Pseudo-logits}_{\text{CLIP}} + (1-\gamma) \cdot \text{TokenFusion}(x, W^*_{\text{LLM}})\}$$

$$\mathcal{L}_{\text{st}} = -\mathbb{E}_{x \in \mathcal{X}_t} \sum_{j=1}^{C} \mathbb{I}\{\hat{y} = j\} log\left(\text{TF}(\mathcal{A}(x), W^*_{\text{LLM}})\right)$$

$$\mathcal{L}_{\text{reg}} = -\frac{1}{C} \sum_{k=1}^{C} \log \bar{p}_{\mathcal{A}(x^k)}$$

$$\mathcal{L} = \mathcal{L}_{\text{st}} + \mathcal{L}_{\text{reg}}$$

$f(x_i)$ 表示第 i 个区域的 CLIP 嵌入

$f(x) = P_{\text{CLIP}}(v^{\text{CLS}})$ 为全局图像表示

将其与固定的细粒度文本分类器 W_{LLM} 对齐，以克服 [CLS] token 的粗粒度限制，并生成一致的伪预测。

Result

Method	Venue	Birdsnap	Caltech	Cars	CIFAR100	DTD	FGVC	Flowers	Food101	Imagenet	Pets	RESISC	SUN397	UCF101	Avg
Zero-shot / Training-free Methods															
CLIP [39]	ICML'21	37.45	90.69	58.70	64.47	44.63	19.50	66.42	83.95	63.30	87.50	57.59	61.32	61.86	61.34
CuPL [38]	ICCV'23	37.02	94.62	60.79	65.22	50.11	20.94	69.51	84.05	64.26	87.16	61.14	65.57	66.90	63.64
WCA* [24]	ICML'24	37.63	94.02	61.95	51.78	51.60	21.15	68.70	83.97	**65.01**	86.32	62.56	64.93	65.82	62.73
UA Methods															
UPL [15]	-	32.80	92.36	49.41	67.41	45.37	17.07	67.40	84.25	58.22	83.84	57.63	62.12	62.04	59.99
POUF [47]	ICML'23	38.40	94.10	57.70	62.00	46.10	18.20	67.80	82.10	52.20	87.80	66.40	60.00	61.20	61.08
LaFTer [33]	NeurIPS'23	31.14	94.39	57.44	69.79	50.32	19.86	72.43	82.45	61.63	84.93	61.60	65.87	65.08	62.07
ReCLIP† [14]	WACV'24	37.38	93.84	58.84	71.43	53.88	18.87	72.63	84.22	63.95	85.27	73.05	65.23	67.06	64.69
DPA‡ [2]	WACV'25	31.54	**95.54**	56.83	74.22	55.96	20.10	75.48	84.76	64.64	90.11	71.11	68.13	66.69	65.78
microCLIP (Ours)	-	**38.59**	94.93	**65.81**	**77.41**	**60.00**	**22.74**	**75.84**	**85.58**	64.45	**90.24**	**77.25**	**68.98**	**70.98**	**68.68**

表 1. 13 个数据集上基于 ViT-B/32 主干网络的最先进方法的 Top-1 准确率（%）对比。* 表示使用与 microCLIP 相同数量的图像块复现的结果。† 我们通过在归纳情景下训练 ReCLIP [14] 得到结果。‡ 为保证公平比较，我们使用与 microCLIP 相同的固定学习率复现 DPA。