Unleashing the Potential of Vision-Language Models for Long-Tailed Visual Recognition

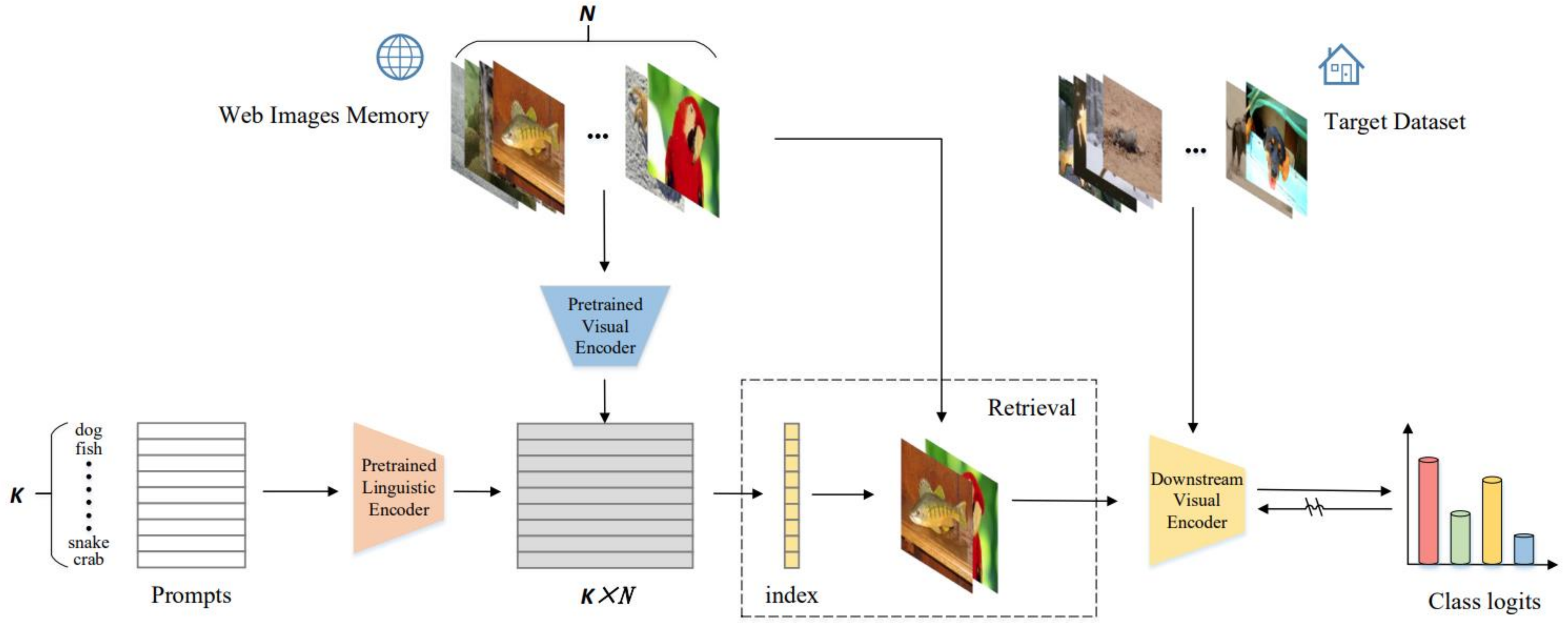
Teli Ma[1], Shijie Geng[2], Mengmeng Wang[3], Sheng Xu[4], Hongsheng Li[1,5], Baochang Zhang[4], Peng Gao[1], Yu Qiao[1]

[1]Shanghai AI Laboratory, [2]Rutgers University, [3]Zhejiang University, [4]Beihang University, [5]MMLab CUHK

Motivation

- All previous Long-tailed recognition models are confined to a predetermined manner which designs models entirely relying on the **visual modality.**
- Explore whether **language modality** can be effective and complementary information for this task.

TACKLE (TrAnsfer Conceptual Knowledge from Language to imagE)



Problem of BALLAD:

- *Contrastive fine-tuning consumes huge computational overheads*

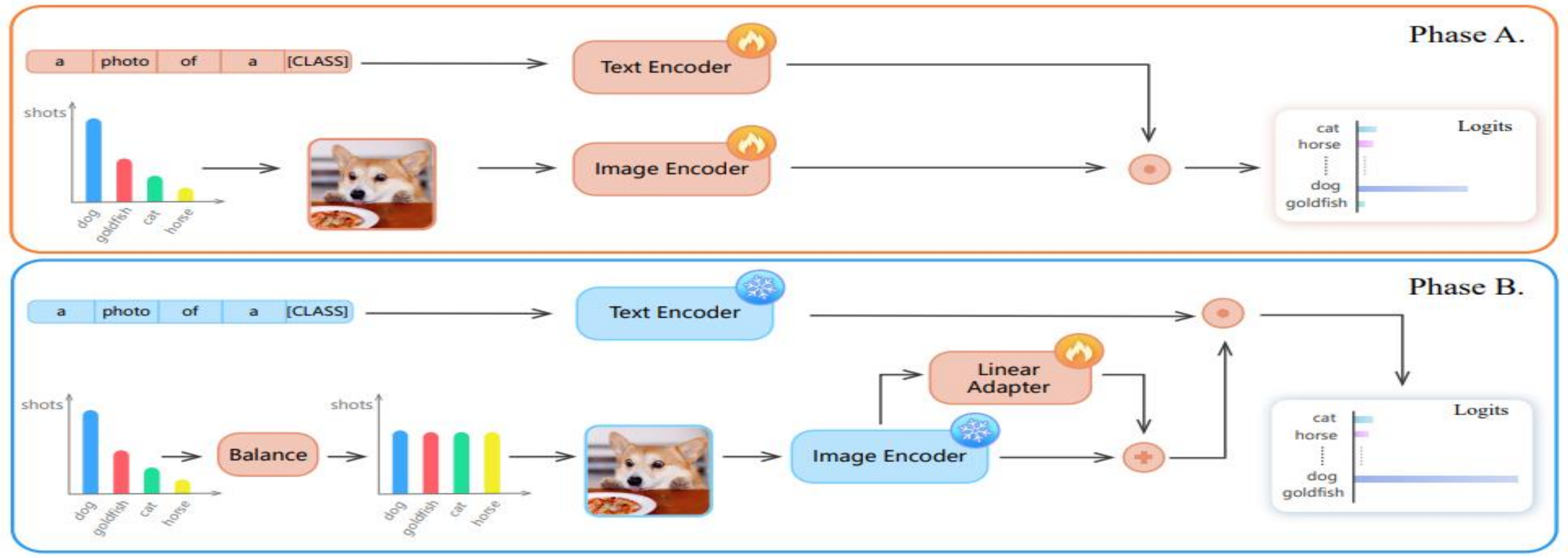
TACKLE:

- *Leveraging conceptual knowledge to generate images in an annotation-free and non-parametric manner.*

Construct dataset for tailed classes from web images based on the probability:

$$p_k = \frac{\exp\left(f_{\mathcal{P}(k)} f_{\mathcal{I}}^{\top}\right)/\tau}{\sum_{j=1}^{K} \exp\left(f_{\mathcal{P}(j)} f_{\mathcal{I}}^{\top}\right)/\tau},$$

BALanced Linear ADapter (BALLAD)



1. Phase A: Contrastive Fine-Tuning

$$\mathcal{L} = \mathcal{L}_{v\to l} + \mathcal{L}_{l\to v}$$

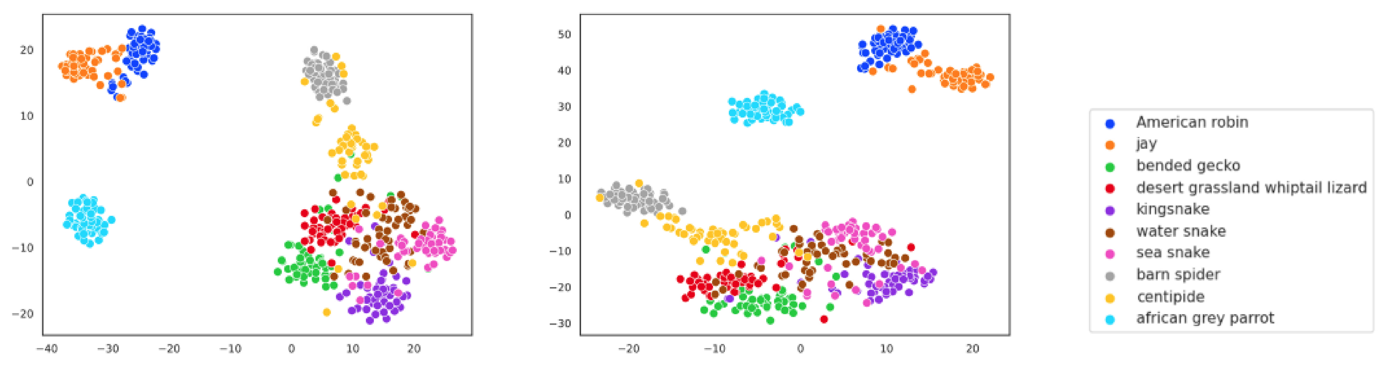
$$= -\frac{1}{|\mathcal{T}_i^{+}|}\sum_{T_j\in\mathcal{T}_i^{+}} \log \frac{\exp\left(\mathbf{v}_i^{\top}\mathbf{u}_j/\tau\right)}{\sum_{T_k\in\mathcal{T}}\exp\left(\mathbf{v}_i^{\top}\mathbf{u}_k/\tau\right)} - \frac{1}{|\mathcal{I}_i^{+}|}\sum_{I_i\in\mathcal{I}_i^{+}} \log \frac{\exp\left(\mathbf{u}_i^{\top}\mathbf{v}_j/\tau\right)}{\sum_{I_k\in\mathcal{I}}\exp\left(\mathbf{u}_i^{\top}\mathbf{v}_k/\tau\right)},$$

2. Phase B: Balanced Adapting

$$\boldsymbol{f}^{\star} = \lambda \cdot \mathrm{ReLU}\left(\mathbf{W}^{\top}\boldsymbol{f} + \mathbf{b}\right) + (1-\lambda)\cdot \boldsymbol{f},$$

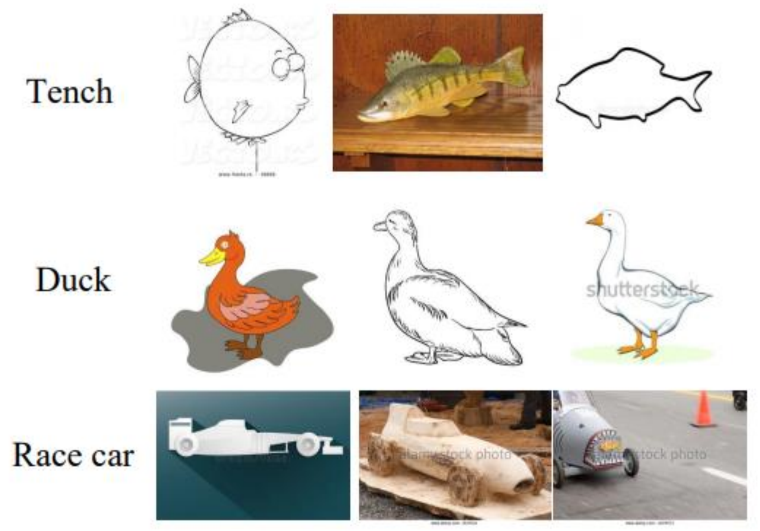
Qualitative Results

✓ **T-SNE visualization**



(a) Decoupled Training (b) Joint Training (c) Legend

✓ **Web images retrieved by TACKLE**



Experiments

Method	ImageNet-LT Backbone	overall
τ-normalized [16]	RN50	46.7
LWS [16]	RN50	47.7
Blanced Softmax [28]	RN50	55.0
RIDE [57]	RN50	55.4
PaCo [illegible]	RN50	57.0
τ-normalized [16]	RN50*	51.3
LWS [16]	RN50*	52.1
PaCo [illegible]	RN50*	60.2
BALLAD	RN50*	67.2
τ-normalized [16]	RX50	49.4
LWS [16]	RX50	49.9
ResLT [3]	RX50	52.9
Blanced Softmax [28]	RX50	56.2
RIDE [57]	RX50	56.8
PaCo [illegible]	RX50	58.2
TACKLE	RX50	60.6
TADE [illegible]	RX50	58.8
τ-normalized [16]	RX101	49.6
LWS [16]	RX101	50.1
ResLT [3]	RX101	55.1
Blanced Softmax [28]	RX101	58.0
PaCo [illegible]	RX101	60.0
TACKLE	RX101	62.5
BALLAD	RN101*	70.5
BALLAD	V-B/16*	75.7
BALLAD	RN50×16*	76.5

Table 1: Long-tailed recognition accuracy on ImageNet-LT for different methods and backbones. ∗ means initializing visual encoder with pretrained weights of CLIP.

Method	Places-LT Backbone	overall
OLTR [22]	RN152♠	35.9
cRT [16]	RN152♠	36.7
τ-normalized [16]	RN152♠	37.9
LWS [16]	RN152♠	37.6
Blanced Softmax [28]	RN152♠	38.6
ResLT [3]	RN152♠	39.8
PaCo [illegible]	RN152♠	41.2
TACKLE	RN152♠	42.6
BALLAD	RN50*	46.5
BALLAD	RN101*	47.9
BALLAD	V-B/16*	49.5
BALLAD	RN50×16*	49.3

Table 2: Long-tailed recognition accuracy on Places-LT for different methods.

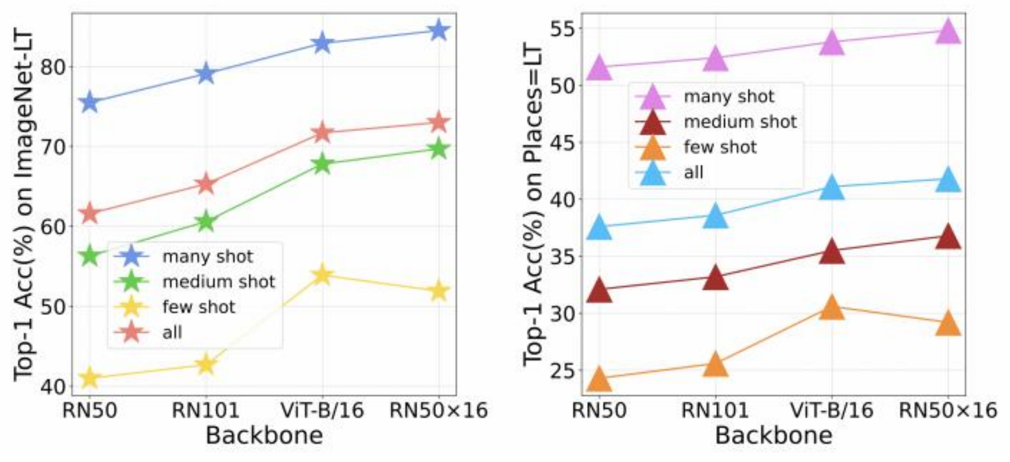
Method	iNaturalist-2018 Backbone	Accuracy(%)
OLTR [22]	RN50	63.9
LWS [16]	RN50	65.9
cRT [16]	RN50	67.6
τ-normalized [16]	RN50	69.3
LADE [illegible]	RN50	69.3
RIDE (2 experts) [57]	RN50	71.4
ResLT [3]	RN50	72.3
RIDE (4 experts) [57]	RN50	72.6
TADE [illegible]	RN50	72.9
PaCo [illegible]	RN50	73.2
TACKLE	RN50	74.4
PaCo [illegible]	RN50*	73.8
BALLAD	RN50*	74.2

Table 3: Long-tailed recognition accuracy on iNaturalist-2018 for different methods. ∗ means initializing visual encoder with pretrained weights of CLIP.

✓ **State-of-the-art on ImageNet-LT(Table 1)**

✓ **State-of-the-art on Places-LT(Table 2)**

✓ **State-of-the-art on iNaturalist-2018 (Table 3)**



(a) ImageNet-LT (b) Places-LT.

✓ **Effectiveness of BALLAD backbones**

Backbones	CLIP	many	medium	few	overall
DeiT-S [34]	-	73.2	59.3	52.3	63.7
CTN-L [7]	-	78.5	62.8	50.2	67.1
DeiT-S [34]	✓	74.3	62.8	58.1	66.6
CTN-L [7]	✓	**78.8**	**65.5**	**55.9**	**69.3**

✓ **Ensemble TACKLE and CLIP**

Visual Backbone	ImageNet-LT zero-shot	ImageNet-LT BALLAD	Places-LT zero-shot	Places-LT BALLAD	iNaturalist-2018 zero-shot	iNaturalist-2018 BALLAD
ResNet-50	58.2	67.2 (+9)	35.3	46.5 (+11.2)	2.6	74.2 (+71.6)
ResNet-101	61.2	70.5 (+9.3)	36.2	47.9 (+11.7)	-	-
ViT-B/16	66.7	75.7 (+9)	37.8	49.5 (+11.7)	-	-
ResNet-50×16	69.0	76.5 (+7.5)	37.1	49.3 (+12.2)	-	-

✓ **Comparison of CLIP zero-shot and BALLAD-Training**