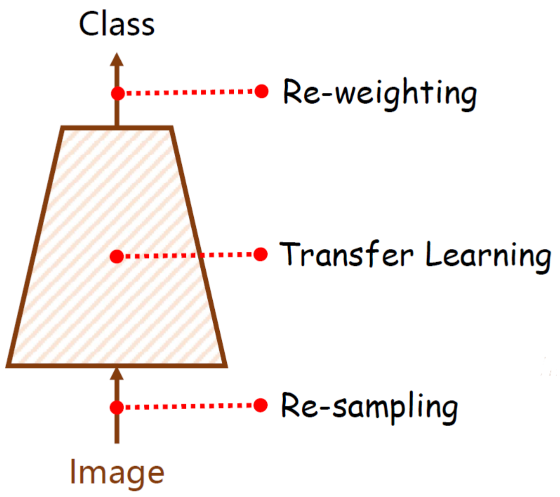
VL-LTR: Learning Class-wise Visual-Linguistic Representation for Long-Tailed Visual Recognition

Changyao Tian[1*†], Wenhai Wang[3*], Xizhou Zhu[2*], Jifeng Dai[2✉], Yu Qiao[3]
[1]Chinese University of Hong Kong [2]SenseTime [3]Shanghai AI Laboratory
tcyhost@buaa.edu.cn {wangwenhai, qiaoyu}@pjlab.org.cn
{zhuwalter, daijifeng}@sensetime.com

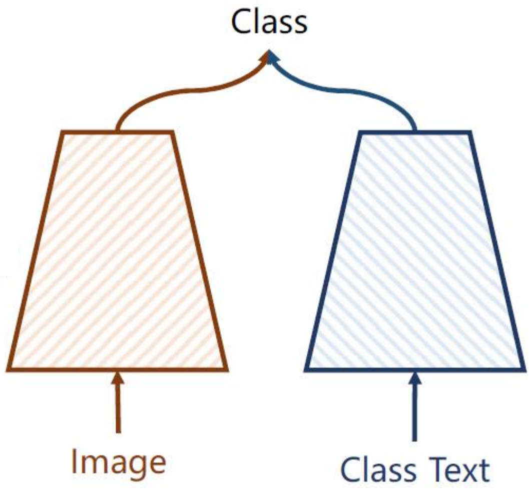
ECCV 2022

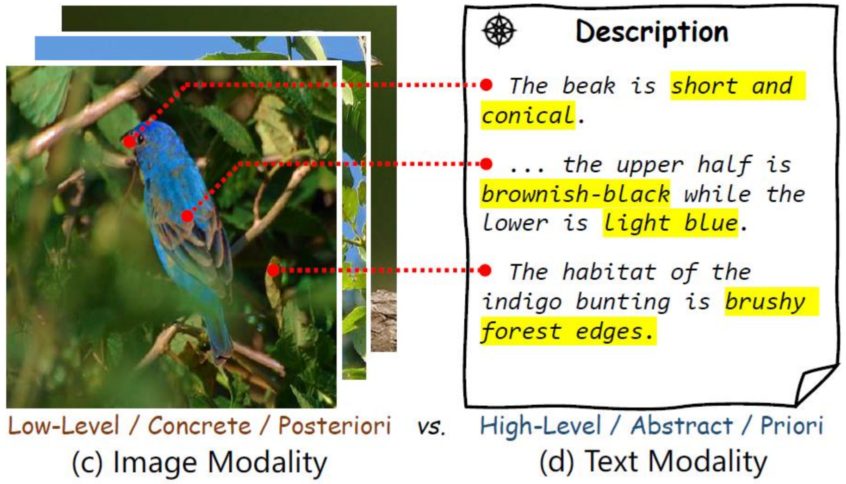
Background



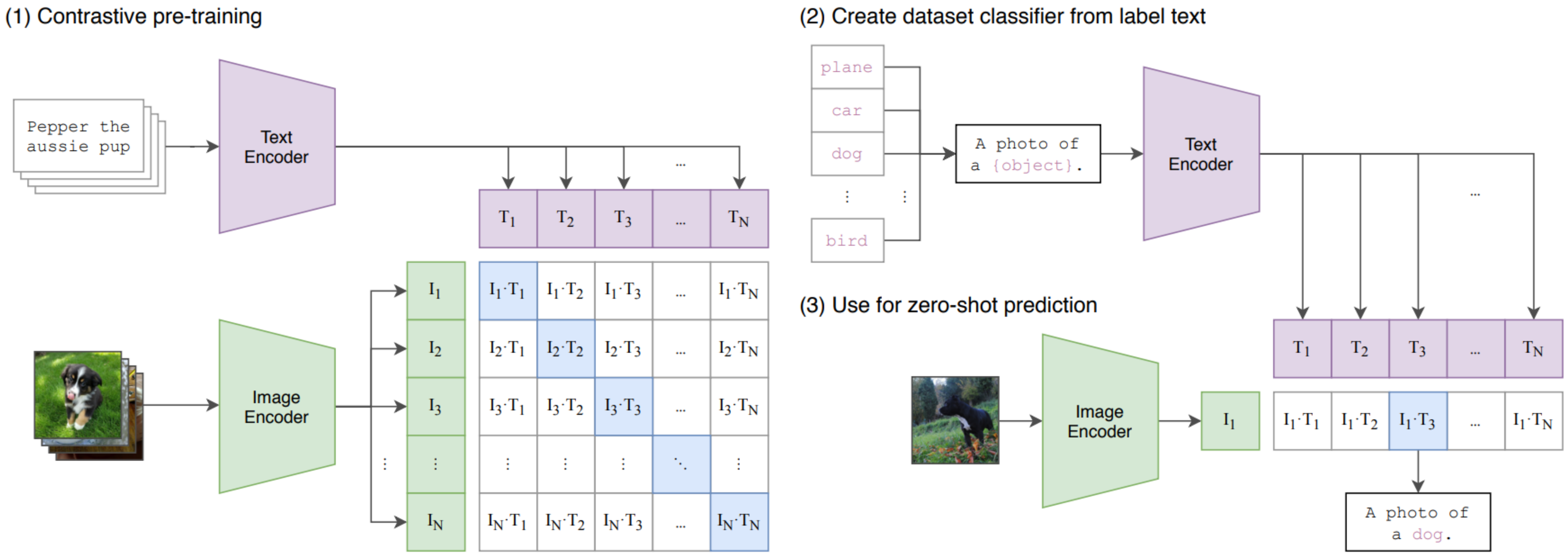
- Re-sampling the training data
- Reweighting the loss functions
- Employing transfer learning methods

There are some inner connections between images and text descriptions of the same class, especially when it comes to some visual concepts and attributes. Text descriptions are prior knowledge that can be summarized by experts, which could be useful when there are no sufficient images to learn general class-wise representation for recognition.



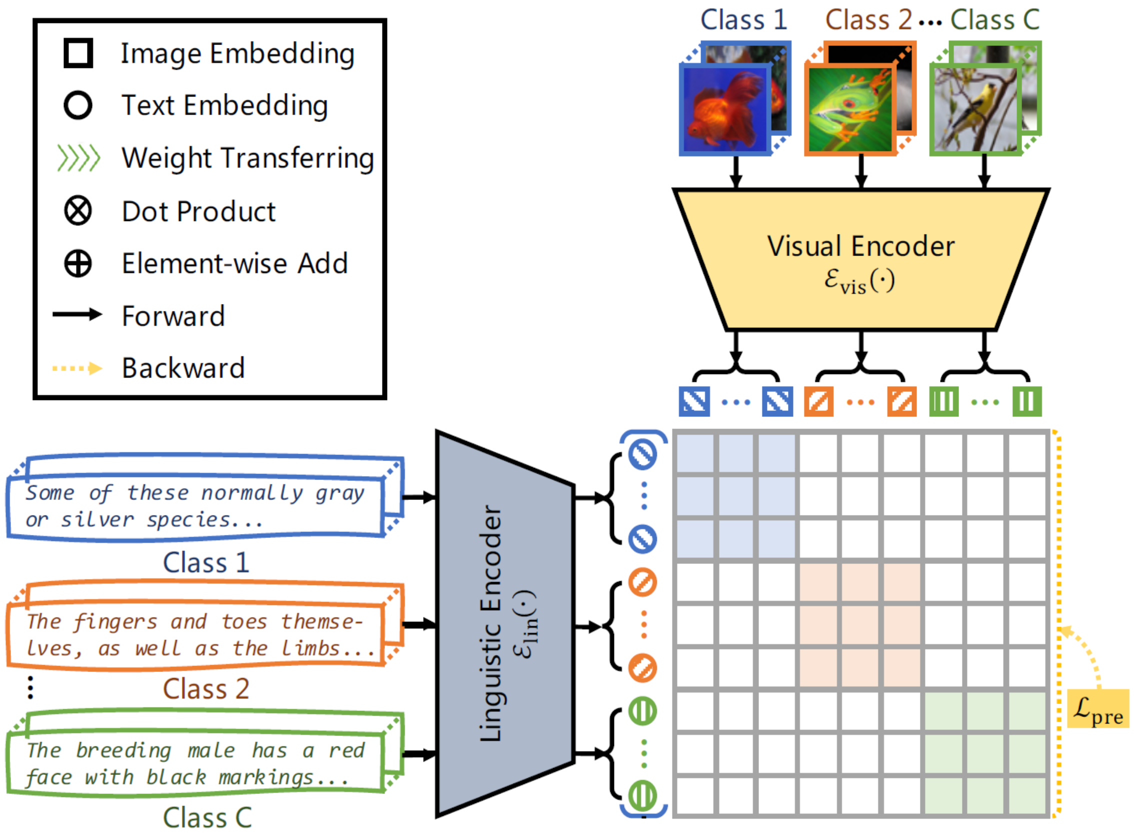


Background


(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
(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
I_1
T_1 T_2 T_3 ... T_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

Methods

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS



$$\boxed{\mathcal{L}_{\text{pre}} = \lambda \mathcal{L}_{\text{ccl}} + (1-\lambda)\mathcal{L}_{\text{dis}}}$$

images $\mathcal{I} = \{I_i\}_{i=1}^N$

text sentences $\mathcal{T} = \{T_i\}_{i=1}^N$

Feed into the visual encoder and linguistic encoder

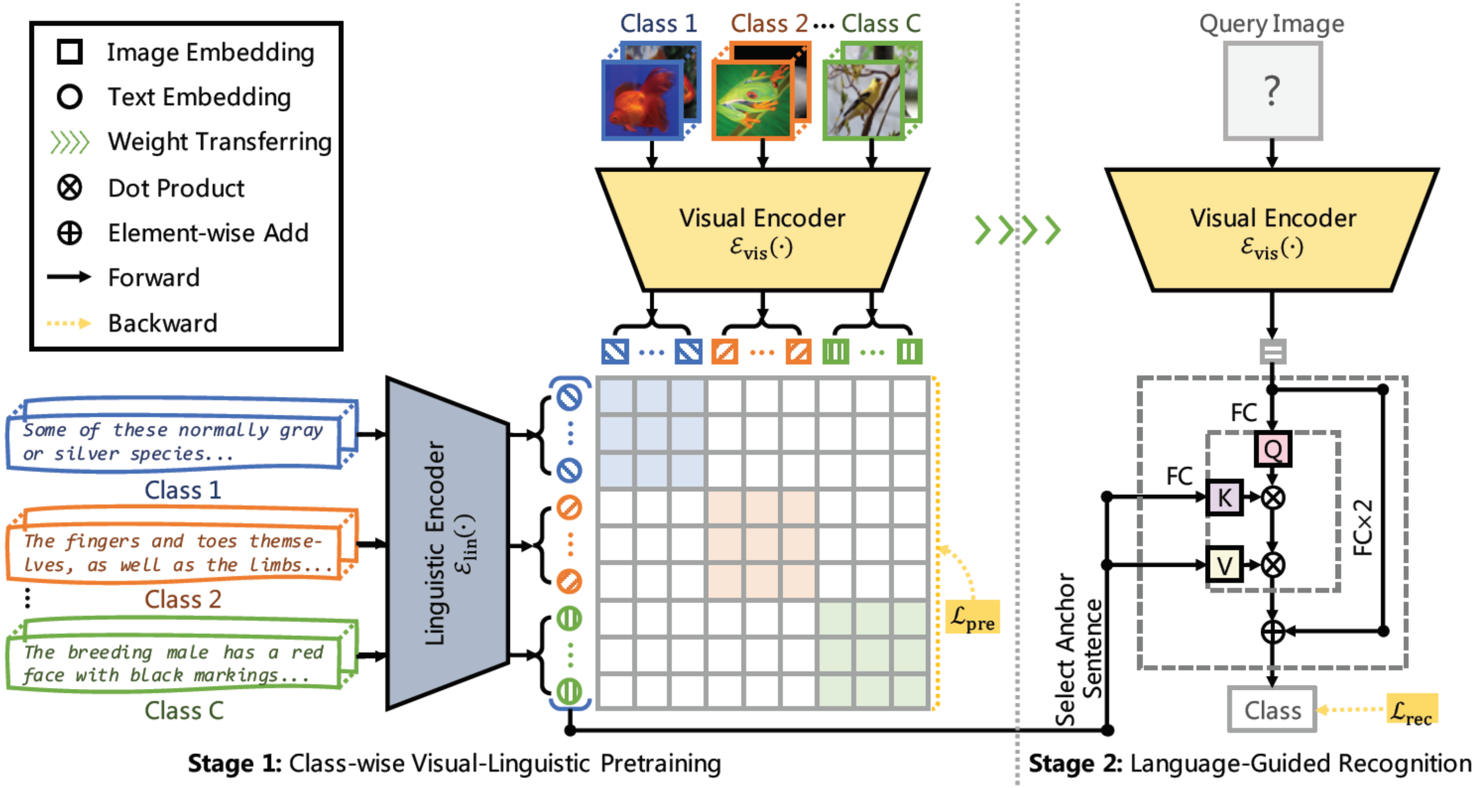
$$E_i^I = \mathcal{E}_{\text{vis}}(I_i), \quad E_i^T = \mathcal{E}_{\text{lin}}(T_i),$$

cosine similarity of E_i^I and E_j^T

$$\begin{aligned}\mathcal{L}_{\text{ccl}} =& \mathcal{L}_{\text{vis}} + \mathcal{L}_{\text{lin}} \\ =& -\frac{1}{|\mathcal{T}_i^+|}\sum_{T_j \in \mathcal{T}_i^+} \log \frac{\exp(\boxed{S_{i,j}}/\tau)}{\sum_{T_k \in \mathcal{T}} \exp(S_{i,k}/\tau)} \\ & -\frac{1}{|\mathcal{I}_i^+|}\sum_{I_j \in \mathcal{I}_i^+} \log \frac{\exp(S_{j,i}/\tau)}{\sum_{I_k \in \mathcal{I}} \exp(S_{k,i}/\tau)},\end{aligned}$$

S' is the cosine similarity matrix produced by the frozen CLIP model

$$\begin{aligned}\mathcal{L}_{\text{dis}} = & -\frac{\exp(\boxed{S'_{i,i}}/\tau)}{\sum_{T_j \in \mathcal{T}} \exp(S'_{i,j}/\tau)} \log \frac{\exp(S_{i,i}/\tau)}{\sum_{T_k \in \mathcal{T}} \exp(S_{i,k}/\tau)} \\ & -\frac{\exp(S'_{i,i}/\tau)}{\sum_{I_j \in \mathcal{I}} \exp(S'_{j,i}/\tau)} \log \frac{\exp(S_{i,i}/\tau)}{\sum_{I_k \in \mathcal{I}} \exp(S_{k,i}/\tau)}.\end{aligned}$$



$$\boxed{\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{CE}}(P^I, \mathbf{y}) + \mathcal{L}_{\text{CE}}(P^T, \mathbf{y})}$$

Anchor Sentence Selection(AnSS):

For each text sentence T_i, we score each sentence T_i, by computing the $\mathcal{L}_{lin}$ between the sentence and the image batch I'. Then, we select M text sentences with the smallest $\mathcal{L}_{lin}$ as the anchor sentences for the follow-up visual recognition.

Language-Guided Recognition Head((LGR)):

$$Q = \text{Linear}(\text{LayerNorm}(E^I)),$$

$$K = \text{Linear}(\text{LayerNorm}(E^T)), \quad V = E^T,$$

$$G = \sigma(\frac{QK^{\top}}{\sqrt{D}})V,$$

the cosine similarity of E^I and G

$$P = \boxed{P^I + P^T} = \sigma(\text{MLP}(E^I)) + \sigma(\boxed{\langle E^I, G \rangle}/\tau)$$

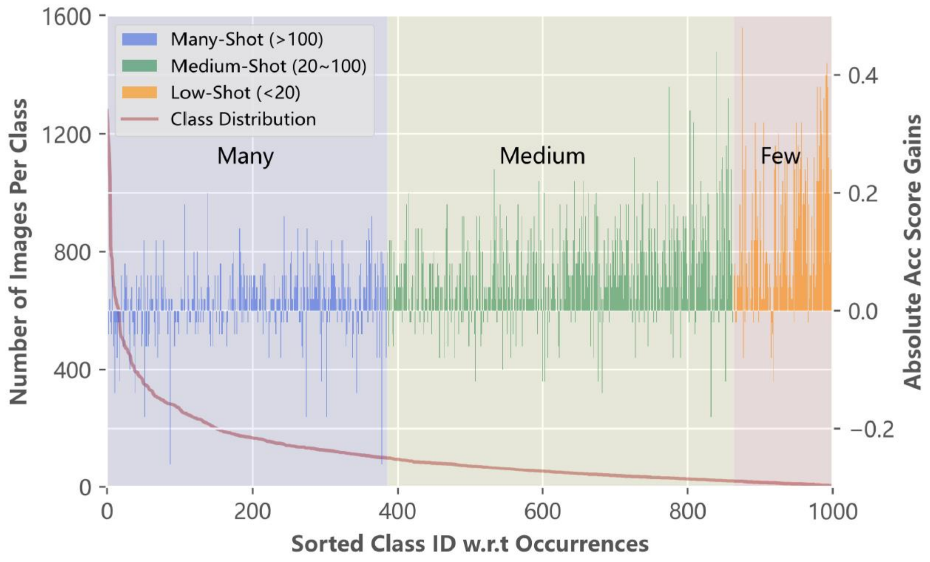
the classification probabilities based on visual and linguistic representation

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

ImageNet-LT:

Method	Backbone	Accuracy (%)			
		Overall	Many	Medium	Few
Cross Entropy [26]	ResNeXt-50	44.4	65.9	37.5	7.7
OLTR [29]	ResNeXt-50	46.3	-	-	-
SSD [26]	ResNeXt-50	56.0	66.8	53.1	35.4
RIDE (4 Experts) [48]	ResNeXt-50	56.8	68.2	53.8	36.0
TADE [53]	ResNeXt-50	58.8	66.5	57.0	43.5
smDRAGON [39]	ResNeXt-50	50.1	-	-	-
ResLT [6]	ResNeXt-101	55.1	63.3	53.3	40.3
PaCo [7]	ResNeXt-101	60.0	68.2	58.7	41.0
NCM [21]	ResNeXt-152	51.3	60.3	49.0	33.6
cRT [21]	ResNeXt-152	52.4	64.7	49.1	29.4
τ-normalized [21]	ResNeXt-152	52.8	62.2	50.1	35.8
LWS [21]	ResNeXt-152	53.3	63.5	50.4	34.2
NCM [21]	ResNet-50*	49.2	58.9	46.6	31.1
cRT [21]	ResNet-50*	50.8	63.3	47.2	27.8
τ-normalized [21]	ResNet50*	51.2	60.9	48.4	33.8
LWS [21]	ResNet-50*	51.5	62.2	48.6	31.8
Zero-Shot CLIP [37]	ResNet-50*	59.8	60.8	59.3	58.6
Baseline	ResNet-50*	60.5	74.4	56.9	34.5
VL-LTR (ours)	ResNet-50*	**70.1**	**77.8**	**67.0**	**50.8**
VL-LTR (ours)	ViT-Base*	**77.2**	**84.5**	**74.6**	**59.3**



Experiments

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

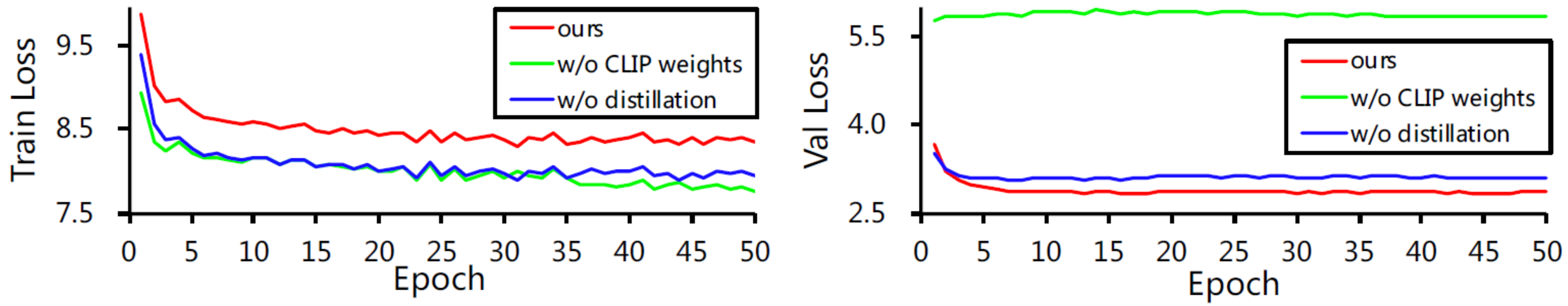
Places-LT:

Method	Backbone	Accuracy (%)			
		Overall	Many	Medium	Few
OLTR [29]	ResNet-152	35.9	44.7	37.0	25.3
ResLT [6]	ResNet-152	39.8	39.8	43.6	31.4
TADE [53]	ResNet-152	40.9	40.4	43.2	36.8
PaCo [7]	ResNet-152	41.2	36.1	47.9	35.3
NCM [21]	ResNet-152	36.4	40.4	37.1	27.3
cRT [21]	ResNet-152	36.7	42.0	37.6	24.9
τ-normalized [21]	ResNet-152	37.9	37.8	40.7	31.8
LWS [21]	ResNet-152	37.6	40.6	39.1	28.6
smDRAGON [39]	ResNet-50	38.1	-	-	-
NCM [21]	ResNet-50*	30.8	37.1	30.6	19.9
cRT [21]	ResNet-50*	30.5	38.5	29.7	17.6
τ-normalized [21]	ResNet-50*	31.0	34.5	31.4	23.6
LWS [21]	ResNet-50*	31.3	36.0	32.1	20.7
Zero-Shot CLIP [37]	ResNet-50*	38.0	37.5	37.5	40.1
Baseline	ResNet-50*	39.7	50.8	38.6	22.7
VL-LTR (ours)	ResNet-50*	**48.0**	**51.9**	**47.2**	**38.4**
VL-LTR (ours)	ViT-Base*	**50.1**	**54.2**	**48.5**	**42.0**

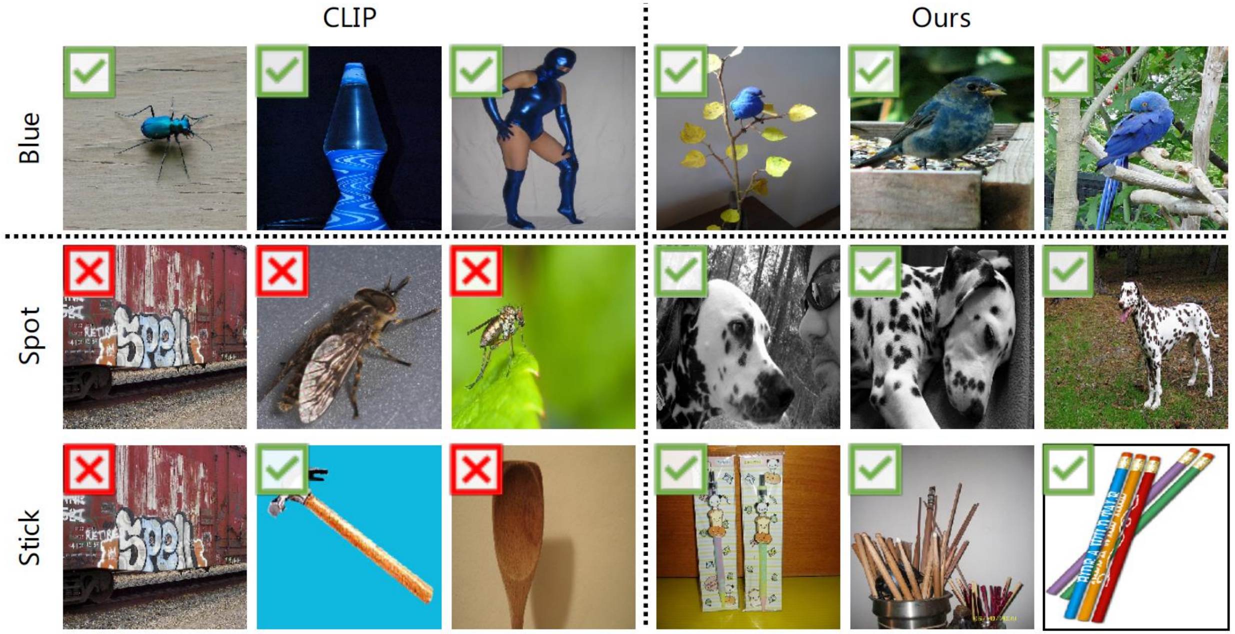
iNaturalist 2018:

Method	Backbone	Accuracy (%)
CB-Focal [2]	ResNet-50	61.1
LDAM+DRW [2]	ResNet-50	68.0
BBN [56]	ResNet-50	69.6
SSD [26]	ResNet-50	71.5
RIDE (4 experts) [48]	ResNet-50	72.6
smDRAGON [39]	ResNet-50	69.1
ResLT [6]	ResNet-50	72.3
TADE [53]	ResNet-50	72.9
PaCo [7]	ResNet-50	73.2
NCM [21]	ResNet-50	63.1
cRT [21]	ResNet-50	67.6
τ-normalized [21]	ResNet-50	69.3
LWS [21]	ResNet-50	69.5
NCM [21]	ResNet-50*	65.3
cRT [21]	ResNet-50*	69.9
τ-normalized [21]	ResNet-50*	71.2
LWS [21]	ResNet-50*	71.0
Zero-Shot CLIP [37]	ResNet-50*	3.4
Baseline	ResNet-50*	72.6
VL-LTR (ours)	ResNet-50*	**74.6**
PaCo [7]	ResNet-152	75.2
DeiT-B/16 [45]	-	73.2
DeiT-B/16-384 [45]	-	79.5
VL-LTR (ours)	ViT-Base*	**76.8**
VL-LTR-384 (ours)	ViT-Base*	**81.0**

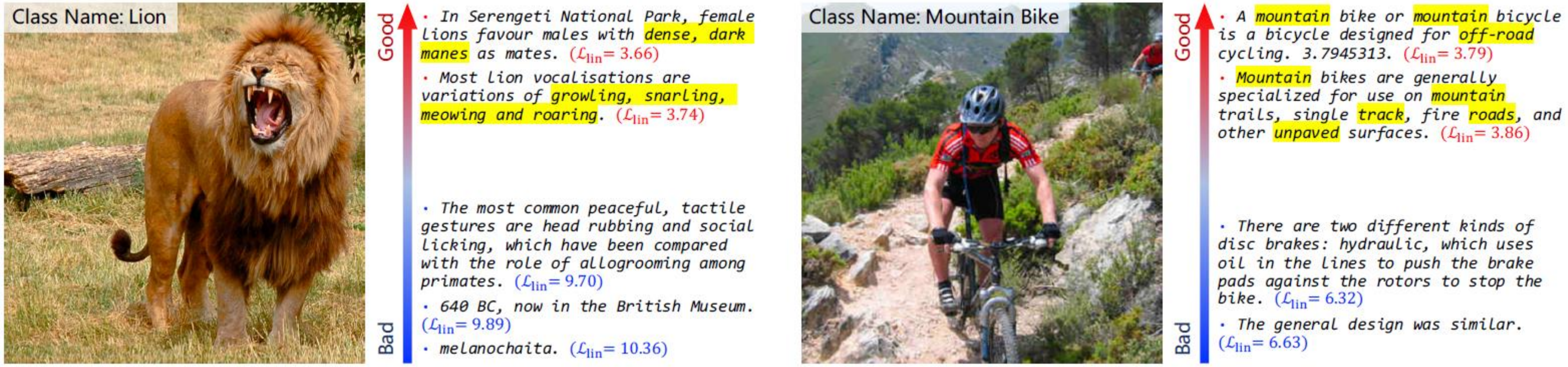
Experiments



#	CLIP Weights	Pre-training w/o $\mathcal{L}_{dis}$	Pre-training w/ $\mathcal{L}_{dis}$	Fine-tuning Head	Fine-tuning SS	Accuracy (%)
1	✓	-	✓	LGR	AnSS	**70.1**
2	✓	-	-	LGR	AnSS	62.8
3	-	✓	-	LGR	AnSS	46.8
4	✓	✓	-	LGR	AnSS	66.2
5	✓	-	✓	FC	-	62.1
6	✓	-	✓	KNN	-	63.9
7	✓	-	✓	LGR	Cut Off	69.7



The method can effectively learn common visual concepts, and even the rare concepts where CLIP makes mistakes, such as "spot" texture and "stick" shape.



Thanks