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

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

A1 Appendices

A Methodology Details

For convenience, we summarize all the notations used in the paper in Table A1.

Table A1: **Summary of notations used in the paper.**

Notation	Meaning
$\mathcal{I} = \{I_i\}_{i=1}^N$	A batch of N image samples
$\mathcal{T} = \{T_i\}_{i=1}^N$	A batch of N text samples
M	Number of anchor sentences per class
$\mathcal{E}_{\text{vis}}(\cdot)$	Visual encoder
$\mathcal{E}_{\text{lin}}(\cdot)$	Linguistic encoder
E_i^I	Embeddings of image I_i
E_i^T	Embeddings of text T_i
$S_{i,j}$	Cosine similarity of E_i^I and E_j^T
$\langle E^I, G \rangle$	Cosine similarity of E^I and G
$\mathcal{L}_{\text{ccl}}$	Class-wise contrastive loss
$\mathcal{L}_{\text{vis}}$	Class-wise contrastive loss for images
$\mathcal{L}_{\text{lin}}$	Class-wise contrastive loss for texts
$\mathcal{L}_{\text{dis}}$	Distillation loss
$\mathcal{L}_{\text{pre}}$	Pre-training loss
$\mathcal{L}_{\text{rec}}$	Recognition loss
$\mathbf{y}$	Ground truth label

B Class-level Corpus Preparation

As described in Section 4.1, we collect class-level text descriptions from Wikipedia and prompt templates provided in [2]. In Figure A1, we display part of text descriptions collected for ImageNet-LT [1], Places-LT [1], and iNaturalist-2018 [3] datasets. We see

Table A2: **Detailed statistics of the class-level text descriptions for each dataset**, where M_{min}, M_{max}, M_{mean}, and M_{Med} denotes the minimum, maximum, mean, and median number of sentences of classes respectively, and L_{Avg} denotes the average number of tokens per sentence.

Dataset	M_{min}	M_{max}	M_{mean}	M_{Med}	L_{Avg}
ImageNet-LT [1]	1	721	127	89	29
Places-LT [1]	2	610	116	77	29
iNaturalist 2018 [3]	1	1774	33	17	26

that since these texts are all crawled from the Internet, it is inevitable to have some noisy text within them.

In addition, we report detailed statistics of the collected text descriptions in Table A2, where we find that even if all the corpus comes from Wikipedia, the text quantity of different classes varies greatly.

C Computation Overhead

As mentioned in Section 3.1, our VL-LTR is a two-stage framework with two encoders. Nevertheless, we would like to point out that the computational cost of our method is almost the same as the vision-based method, since the linguistic encoder is not necessary at the inference stage. Specifically, after pre-training, the text embeddings of anchor sentences can be pre-populated offline. During inference, we only need to load the pre-populated text embeddings to perform visual recognition. As reported in Table A3, the GFLOPs and the inference speed of our method are similar to the baseline. These results are tested with a batch size of 128 on one V100 GPU and one 2.20GHz CPU in a single thread. Moreover, we believe such conclusion also applies to other backbones such as ViT, Swin, TransFG, and complemental attention, since our framework is orthogonal to the backbone's structure.

Table A3: **Computation overhead comparison of our VL-LTR (ResNet-50) and the baseline (ResNet-50).** Our method has almost the same GFLOPs and inference speed to the baseline. GFLOPs is calculated under the input scale of 224×224.

Method	GFLOPs	Time Cost (ms)
Baseline	5.4	1.1
VL-LTR (ours)	5.5	1.3

D Comparison with Zero-Shot CLIP

In Table A4, we compare our results and the zero-shot results of CLIP [2] on ImageNet-LT [1], Places-LT [1] and iNaturalist 2018 [3] datasets, respectively. We see that the

Table A4: **Comparison with Zero-Shot CLIP**. Our method achieves improvements on all datasets and is robust to datasets of different domains.

Dataset	Method	Accuracy(%)			
		Overall	Many	Medium	Few
ImageNet-LT	Zero-Shot	59.8	60.8	59.3	**58.6**
	Baseline	60.5	74.4	56.9	34.5
	VL-LTR (ours)	**70.1**	**77.8**	**67.0**	50.8
Places-LT	Zero-Shot	38.0	37.5	37.5	**40.1**
	Baseline	39.7	50.8	38.6	22.7
	VL-LTR (ours)	**48.0**	**51.9**	**47.2**	38.4
iNaturalist 2018	Zero-Shot	3.4	6.1	3.3	2.9
	Baseline	72.6	76.6	74.1	70.2
	VL-LTR (ours)	**74.6**	**78.3**	**75.5**	**72.7**

performance of CLIP drops sharply when the domain of target data (*e.g.*, iNaturalist 2018) is inconsistent with its training data, while our method can achieve significant improvement on all datasets.

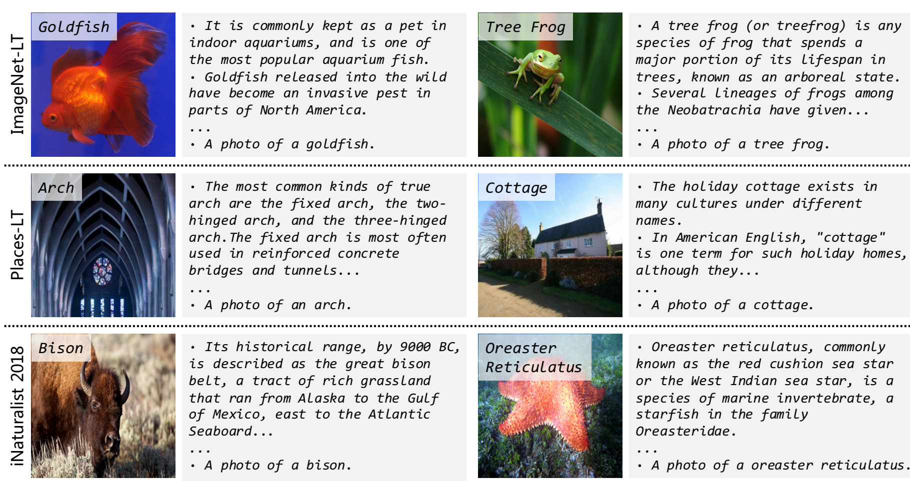

Fig. A1: **Examples of text descriptions crawled from Wikipedia for ImageNet-LT [1], Places-LT [1] and iNaturalist-2018 [3]**, in which both redundant useful and noise information can be found.

E Comparison of Different Distillation Methods in CVLP

To further study the influence of distillation in the pre-training phase, we try to use the pre-trained CLIP model [2] as the teacher model to distill the visual and linguistic encoder of our model at the feature level, in addition to the logits distillation mentioned in Section 3.2. As reported in Table A5, both feature distillation and logits distillation can improve recognition accuracy, and our method achieves the highest accuarcy on ImageNet-LT [1] when using logits distillation with the loss weight λ of 0.5.

Table A5: **Results of different types of distillation in CVLP on ImageNet-LT [1].** Our method achieves the highest accuarcy when using logits distillation with the loss weight λ of 0.5.

Distill Level	λ	Accuracy (%)			
		Overall	Many	Medium	Few
-	0	66.2	76.9	63.5	42.5
Feature	0.1	67.3	77.3	64.4	44.0
	0.5	68.0	77.6	65.2	45.5
Logits	0.1	68.3	**77.9**	65.3	45.1
	0.5 (ours)	**70.1**	77.8	**67.0**	**50.8**

F Comparison of Different Text Description Sources

In Table A6, we compare the results of models using different kinds of text descriptions on ImageNet-LT [1]. Specifically, we use the prompt sentences provided in [2] as the source of text description. We mark this model as "prompt only", and compare it with the default model that uses both Wikipedia and prompt templates as the source of text description (*i.e.*, "wiki + prompt"). We see that "wiki + prompt" outperforms "prompt only" in overall, medium, and few accuracy, which demonstrates the effectiveness of corpus from Wikipedia.

We also notice that although "prompt only" is not the best, its performance is still relatively competitive compared to the vision-based methods (*e.g.*, the strong Baseline established in this work). We attribute this phenomenon to reasons as follows: (1) Our method can make effective use of the pre-trained image and text encoder of CLIP [2], while vision-based methods can only use image encoder; (2) Some class names themselves contain discriminative language information, such as "gold fish", "tree frog", and "mountain bike".

G Visualization of AnSS

To intuitively show the effectiveness of our anchor sentence selection (AnSS), we also present some sentences recommended or filtered out by our AnSS of different classes in Figure A2. We see that our method can reserve useful texts and drop the useless ones effectively.

Table A6: **Results of using different text source on ImageNet-LT [1] and Places-LT [1],** where we see that "wiki + prompt" outperforms "prompt only" in overall, medium, and few accuracy.

Dataset	Source	Accuracy(%)			
		Overall	Many	Medium	Few
ImageNet-LT	Baseline	60.5	74.4	56.9	34.5
	prompt only	69.4	**77.9**	66.5	49.3
	wiki + prompt (ours)	**70.1**	77.8	**67.0**	**50.8**
Places-LT	Baseline	39.7	50.8	38.6	22.7
	prompt only	47.3	**52.7**	46.8	36.3
	wiki + prompt (ours)	**48.0**	51.9	**47.2**	**38.4**

H More Examples of Concept Visualization

In this section, we provide more concept visualization results of VL-LTR (ResNet-50) trained on ImageNet-LT [1]. As shown in Figure A3, our models can not only learn some appearance attributes such as the shape and texture, but also understand high-level concepts like "wall" and "sky". Moreover, benefiting from CVLP, our method can cover more visual concepts than CLIP.

References

1. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2019) 1, 2, 3, 4, 5, 6
2. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021) 1, 2, 3, 4, 7
3. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2018) 1, 2, 3

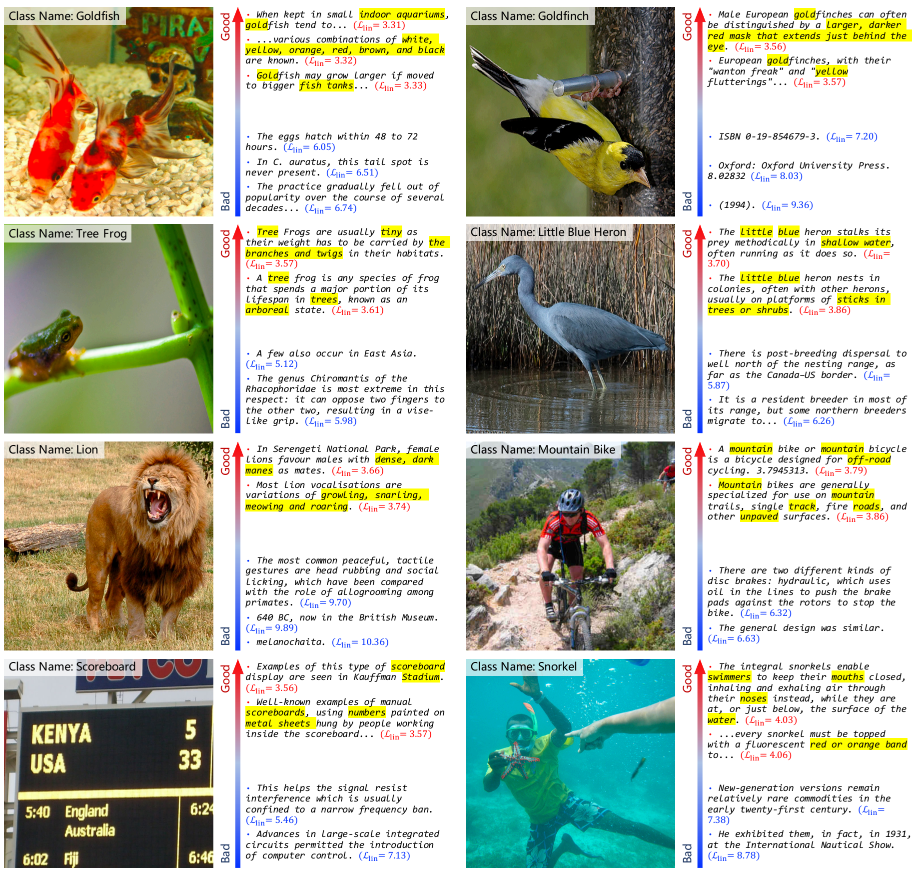

Fig. A2: **Some "good" and "bad" sentences and their corresponding $\mathcal{L}_{\text{lin}}$ of classes in ImageNet-LT [1].** The value of $\mathcal{L}_{\text{lin}}$ can reflect the usefulness of these sentences to some extent, which thereby supports the effectiveness of our AnSS.

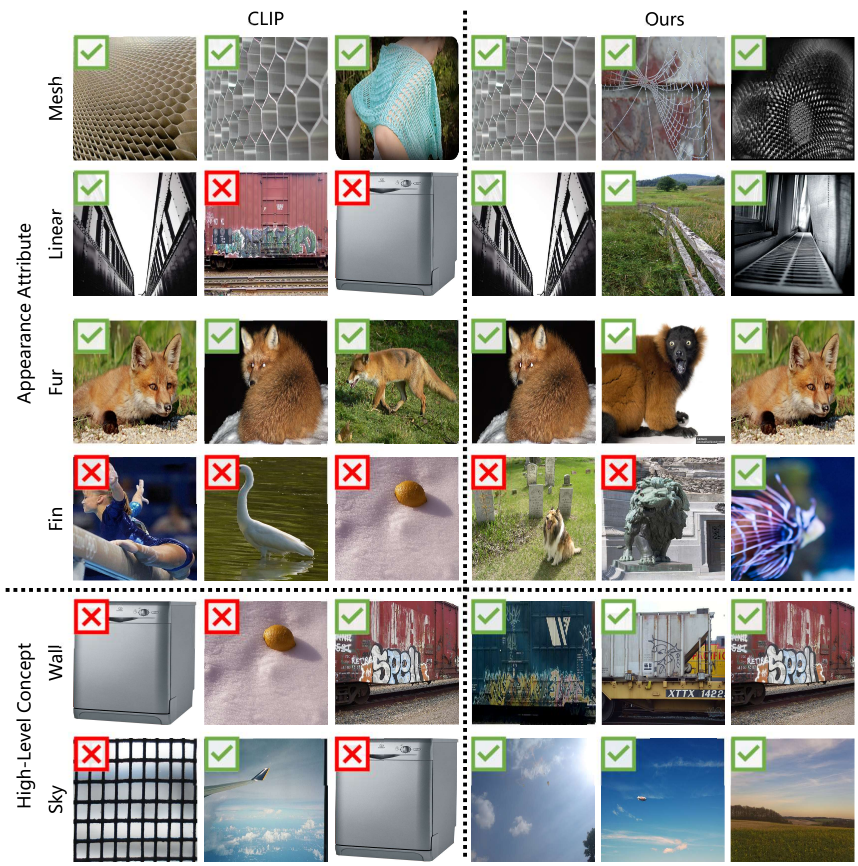

Fig. A3: **Examples of concept visualization.** Our method can not only learn the texture (*e.g.*, mesh) and shape (*e.g.*, linear) of objects, but can also understand some visual attributes (*e.g.*, fur and fin) and high-level concepts (*e.g.*, wall and sky). In addition, compared to the original CLIP [2], our method can cover more visual concepts.