南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Efficient and Long-Tailed Generalization for Pre-trained Vision-Language Model

Jiang-Xin Shi*
National Key Laboratory for Novel Software Technology
School of Artificial Intelligence
Nanjing University, China
shijx@lamda.nju.edu.cn

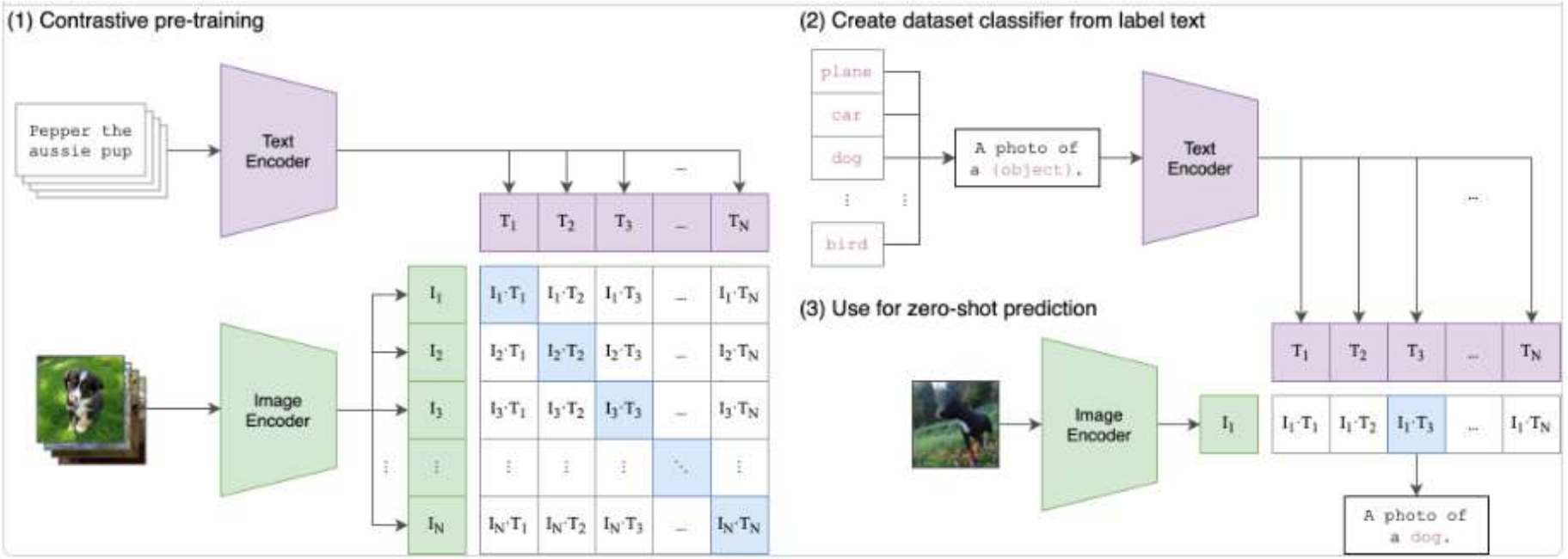
Chi Zhang*
National Key Laboratory for Novel Software Technology
Nanjing University, China
chi-zhang@smail.nju.edu.cn

Tong Wei
School of Computer Science and Engineering
Key Laboratory of Computer Network and Information Integration of Ministry of Education
Southeast University, China
weit@seu.edu.cn

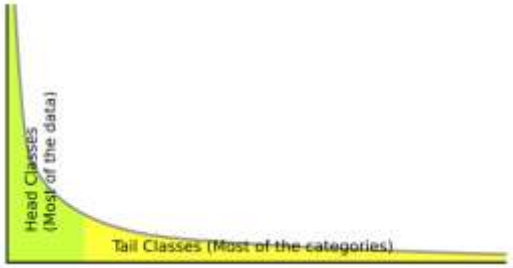
Yu-Feng Li†
National Key Laboratory for Novel Software Technology
School of Artificial Intelligence
Nanjing University, China
liyf@lamda.nju.edu.cn

Background

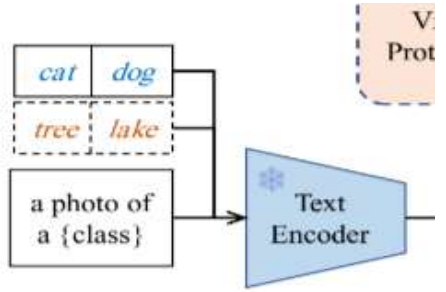
南京航空航天大學
Nanjing University of Aeronautics and Astronautics



Pre-trained vision-language models like CLIP have shown powerful zero-shot inference ability via image-text matching and prove to be strong few-shot learners in various downstream tasks.



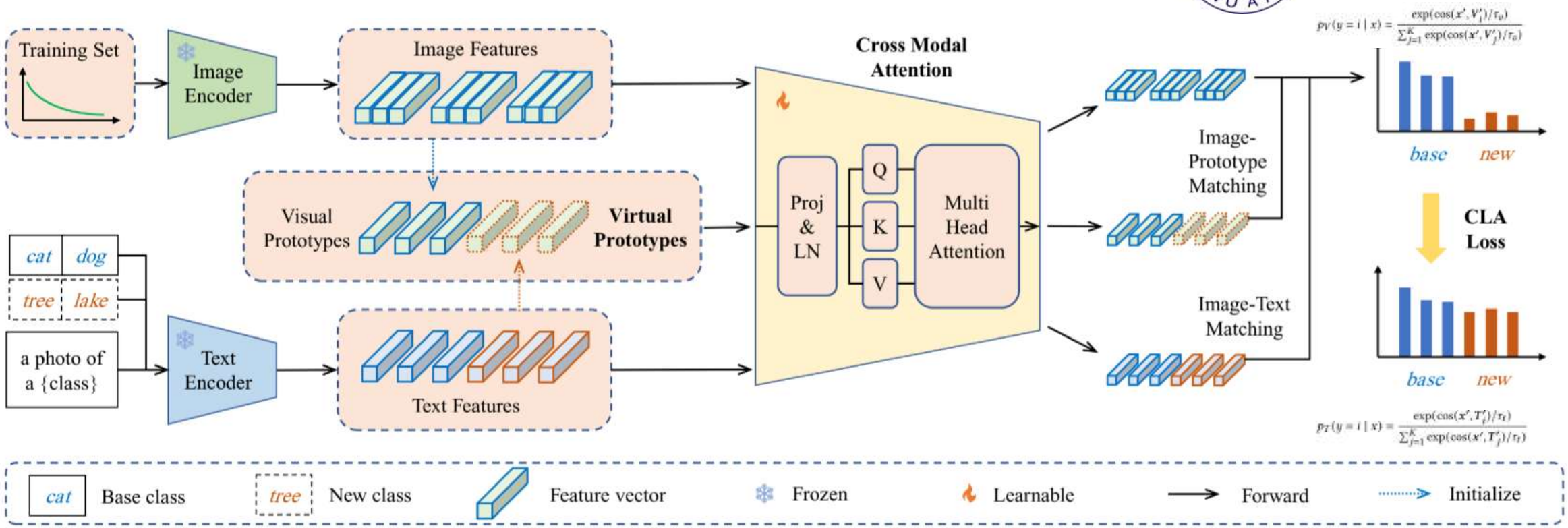
Challenge 1: long-tailed data distribution



Challenge 2: new classes with no samples

Framework

南京航空航天大學
Nanjing University of Aeronautics and Astronautics



Virtual prototypes ▷ New classes with no samples CLA Loss ▷ Long-tailed distribution

$$p(y=i \mid w)=\frac{\exp(\cos(\boldsymbol{x}', \boldsymbol{V}_i')/\tau_v)}{\sum \exp(\frac{\cos(\boldsymbol{x}', \boldsymbol{V}')}{\tau_v}) + \sum \exp(\frac{\cos(\boldsymbol{x}', \hat{\boldsymbol{V}}')}{\tau_v})}$$

$$\mathcal{L}_{cla}(z, y=j) = -\log \frac{\exp(z_j + \log p(y=j))}{\sum_{k=1}^{K} \exp(z_k + \log p(y=k))}$$

Experiments

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Dataset	Classes	Train	Test	Balanced
ImageNet [5]	1000	1.28M	50000	✓
Caltech101 [9]	100	4128	2456	✗
OxfordPets [26]	37	2944	3669	✓
StanfordCars [17]	196	6509	8041	✓
Flowers102 [23]	102	4093	2463	✗
Food101 [1]	101	50500	30300	✓
FGVCAircraft [21]	100	3334	3333	✓
SUN397 [41]	397	15880	19850	✓
DTD [4]	47	2820	1692	✓
EuroSAT [12]	10	13500	8100	✗
UCF101 [34]	101	7639	3783	✗

Dataset	Classes	Test	Balanced
ImageNet-A [14]	200	7500	✗
ImageNetV2 [28]	1000	10000	✓
ImageNet-Sketch [38]	1000	50889	✓
ImageNet-R [13]	200	30000	✗

1. Base-to-new generalization
2. Cross-dataset transfer
3. Domain generalization

调和平均准确率

Table 5: Harmonic mean values of base-to-new accuracy (%) of different methods on datasets with imbalance ratios 10, 20, 50. The models are trained on an imbalanced base set and then evaluated on both base and new classes. Harmonic accuracy is calculated by $\frac{2\times base\times acc}{base+new}$ to highlight the generalization trade-off. The best results are presented in bold.

(a) Imbalance Ratio = 10.

	Cal.	OP.	SC.	FLw.	Food.	FA.	SUN.	DTD.	ES.	UCF.	Avg.
CoOp + LogitAdjusted Loss	91.78	94.10	69.23	71.86	89.25	32.12	72.15	54.88	54.42	64.74	70.66
CoCoOp + LogitAdjusted Loss	95.09	**96.69**	71.91	77.61	**91.20**	33.71	77.99	65.11	60.28	76.78	75.67
Linear Feature Alignment	95.69	94.09	72.72	84.38	90.44	34.27	78.39	67.43	69.56	82.71	76.97
Candle (Ours)	**95.89**	95.99	**74.30**	**85.03**	90.80	**37.78**	**79.26**	**68.13**	**80.51**	**83.17**	**79.34**

(b) Imbalance Ratio = 20.

	Cal.	OP.	SC.	FLw.	Food.	FA.	SUN.	DTD.	ES.	UCF.	Avg.
CoOp + LogitAdjusted Loss	92.65	94.15	67.39	73.72	86.38	29.33	68.93	55.18	62.64	60.12	70.08
CoCoOp + LogitAdjusted Loss	95.25	**96.64**	71.38	80.31	**91.20**	32.78	77.29	61.31	58.82	71.70	74.48
Linear Feature Alignment	95.56	90.90	70.35	84.03	89.72	33.02	77.30	66.07	68.74	81.80	75.75
Candle (Ours)	**95.84**	95.89	**73.49**	**84.92**	90.75	**38.02**	**78.53**	**67.32**	**80.96**	**82.59**	**79.08**

(c) Imbalance Ratio = 50.

	Cal.	OP.	SC.	FLw.	Food.	FA.	SUN.	DTD.	ES.	UCF.	Avg.
CoOp + LogitAdjusted Loss	93.30	93.34	67.18	75.45	87.65	29.20	65.91	51.42	57.35	61.62	69.92
CoCoOp + LogitAdjusted Loss	94.90	95.44	69.97	76.84	**91.10**	31.45	76.18	59.37	64.99	77.53	74.32
Linear Feature Alignment	94.23	86.76	67.95	82.81	87.73	30.75	75.13	61.78	61.91	79.49	72.85
Candle (Ours)	**94.95**	**95.83**	**71.78**	**84.62**	90.70	**36.68**	**78.05**	**65.69**	**80.17**	**81.72**	**78.20**

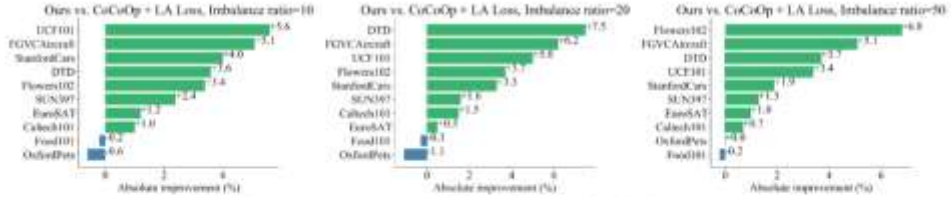


Figure 3: Absolute improvement on the base classes with imbalance ratio 10, 20, 50

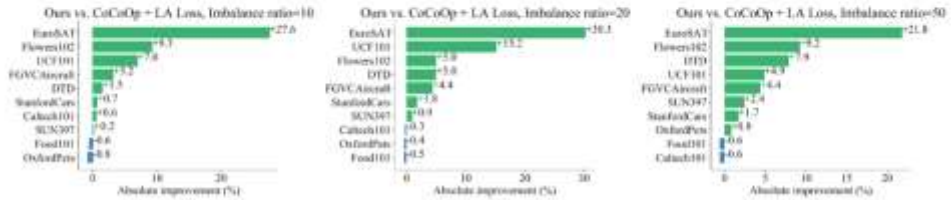


Figure 4: Absolute improvement on the new classes with imbalance ratio 10, 20, 50

Experiments & Ablation study

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Table 6: Comparison of different methods in 16-shot base-to-new generalization. We report the accuracy (%) on both base and new classes, as well as their harmonic mean. The best results are presented in bold.

(a) Average over 11 datasets.

	Base	New	H
CLIP	69.34	74.22	71.70
CoOp	82.69	63.22	71.66
CoCoOp	80.47	71.69	75.83
LFA	83.62	74.56	78.83
Ours	**83.86**	**76.55**	**80.04**

(b) ImageNet.

	Base	New	H
CLIP	72.43	68.14	70.22
CoOp	76.47	67.88	71.92
CoCoOp	75.98	**70.43**	**73.10**
LFA	76.89	69.36	72.93
Ours	**76.97**	68.54	72.48

(c) Caltech101.

	Base	New	H
CLIP	96.84	94.00	95.40
CoOp	98.00	89.91	93.73
CoCoOp	97.96	93.81	95.84
LFA	98.41	93.93	96.13
Ours	**98.54**	**94.47**	**96.46**

(d) OxfordPets.

	Base	New	H
CLIP	91.17	97.26	94.12
CoOp	93.67	95.29	94.47
CoCoOp	95.20	**97.69**	**96.43**
LFA	95.13	96.23	95.68
Ours	**95.53**	97.34	**96.43**

(e) StanfordCars.

	Base	New	H
CLIP	63.37	74.89	68.85
CoOp	78.12	60.40	68.13
CoCoOp	70.49	73.59	72.01
LFA	76.32	74.88	75.59
Ours	**79.14**	**74.92**	**76.97**

(f) Flowers102.

	Base	New	H
CLIP	72.08	**77.80**	74.83
CoOp	97.60	59.67	74.06
CoCoOp	94.87	71.75	81.71
LFA	97.34	75.44	85.00
Ours	**98.01**	77.52	**86.57**

(g) Food101.

	Base	New	H
CLIP	90.10	91.22	90.66
CoOp	88.33	82.26	85.19
CoCoOp	**90.70**	91.29	90.99
LFA	90.52	**91.48**	**91.00**
Ours	90.52	91.23	90.87

(h) FGVCAircraft.

	Base	New	H
CLIP	27.19	36.29	31.09
CoOp	40.44	22.30	28.75
CoCoOp	33.41	23.71	27.74
LFA	41.48	32.29	36.31
Ours	**43.86**	**36.69**	**39.96**

(i) SUN397.

	Base	New	H
CLIP	69.36	75.35	72.23
CoOp	80.60	65.89	72.51
CoCoOp	79.74	76.86	78.27
LFA	**82.13**	77.20	79.59
Ours	81.64	**77.93**	**79.74**

(j) DTD.

	Base	New	H
CLIP	53.24	59.90	56.37
CoOp	79.44	41.18	54.24
CoCoOp	77.01	56.00	64.85
LFA	81.29	60.63	69.46
Ours	**81.40**	**61.35**	**69.97**

(k) EuroSAT.

	Base	New	H
CLIP	56.48	64.05	60.03
CoOp	92.19	54.74	68.90
CoCoOp	87.49	60.04	71.21
LFA	**93.40**	71.24	80.83
Ours	89.97	**81.33**	**85.43**

(l) UCF101.

	Base	New	H
CLIP	70.53	77.50	73.85
CoOp	84.69	56.05	67.46
CoCoOp	82.33	73.45	77.64
LFA	86.97	77.48	81.95
Ours	**87.13**	**80.51**	**83.69**

Table 7: Cross-dataset transfer learning accuracy (%) of different methods. The methods are trained on an imbalanced source dataset (ImageNet) and subsequently evaluated on the target datasets. The best results are presented in bold.

	Cal.	OP.	SC.	FLw.	Food.	FA.	SUN.	DTD.	ES.	UCF.	Avg.
CoOp + LogitAdjusted Loss	90.8	87.0	64.9	67.3	85.3	18.8	63.2	42.2	44.4	65.9	63.0
CoCoOp + LogitAdjusted Loss	**91.4**	88.6	**65.6**	**69.4**	**86.3**	23.0	66.0	**45.0**	42.8	**67.5**	64.6
Candle (Ours)	91.3	**88.9**	64.6	68.3	85.5	**24.2**	**66.1**	44.6	**48.4**	67.2	**64.9**

Table 8: Domain generalization accuracy (%) of different methods. The methods are trained on an imbalanced source dataset (ImageNet) and subsequently evaluated on the target datasets. The best results are presented in bold.

	IN.	IN-A.	INV2.	IN-S	IN-R.
CoOp + LA Loss	70.7	48.7	**63.5**	47.2	73.8
CoCoOp + LA Loss	71.3	**49.1**	63.3	47.8	74.4
Candle (Ours)	**71.6**	**49.1**	62.8	**48.3**	**75.0**

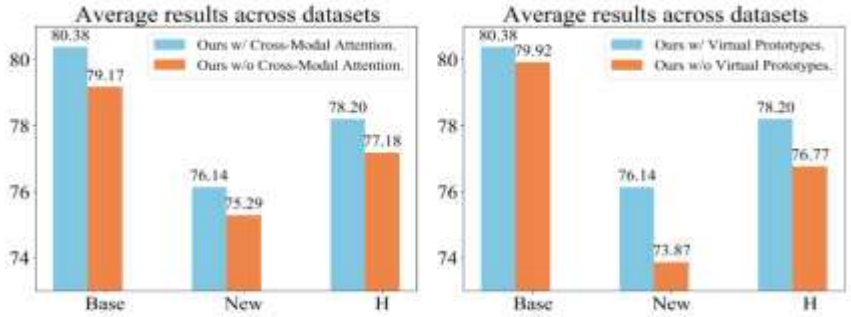


Figure 5: Ablation studies on cross-modal attention (left) and virtual prototypes (right). The experiment is conducted on the imbalanced base-to-new generalization task with an imbalance ratio of 50.

Thanks