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

Teli Ma[1]
telima9868@gmail.com
Shijie Geng[2]
sg1309@rutgers.edu
Mengmeng Wang[3]
mengmengwang@zju.edu.cn
Sheng Xu[4]
shengxu@buaa.edu.cn
Hongsheng Li[1,5]
hsli@ee.cuhk.edu.hk
Baochang Zhang[4]
bczhang@buaa.edu.cn
Peng Gao[1]
gaopeng@pjlab.org.cn
Yu Qiao[1,✉]
qiaoyu@pjlab.org.cn

[1] Shanghai AI Laboratory, Shanghai, China
[2] Rutgers University, New Jersey, US
[3] Zhejiang University, Zhejiang, China
[4] Beihang University, Beijing, China
[5] MMLab, The Chinese University of Hong Kong, Hong Kong

A More Implementation Details

A.1 Web-images Datasets

We conclude the details of source datasets where we retrieve images from in TACKLE. Note that only images are leveraged for training while captions of images are omitted.

Conceptual Captions 3M & 12M. [5, 38] Conceptual Captions 3M (CC3M) is a dataset consisting of 3.3M images annotated with captions. All the images and their raw text descriptions are collected from the web, covering a variety of styles and scenes. The CC3M dataset is programmatically created using a Flume [4]. This pipeline processes billions of Internet webpages in parallel. From these webpages, it extracts, filters, and processes candidate $< image, caption >$ pairs. CC12M is created due to the insight of specific downstream V+L tasks (*e.g.*, VQA, image captioning) can be overly restrictive if the goal is to collect large-scale V+L annotations [5]. Compared with CC3M, CC12M has around 12.4M image-text pairs, about 4× larger than the CC3M. It has a much lower token (word count) to type (vocab

finetuning config	value
optimizer	SGD
visual learning rate	1e-5
linguistic learning rate	1e-5
weight decay	5e-4
momentum	0.9
batch size	512
epochs	50
sampler	instance-aware
learning rate schedule	cosine decay
adapting config	**value**
optimizer	SGD
learning rate	0.2
weight decay	5e-4
momentum	0.9
batch size	2048
epochs	10
sampler	class-aware
adapting ratio	$\lambda = 0.2$
learning rate schedule	cosine decay
Data augmentation	**value**
image size	224
random crop	scale=(0.5,1)
interpolation	BICUBIC
random horizontal flip	p=0.5

Table 1: Training configs of BALLAD on ImageNet-LT.

size) ratio, indicating a longer-tail distribution and a higher diversity degree of the concepts captured. Meanwhile, the average length of descriptions in CC12M is much longer.

SBU Captioned Photo Dataset. [33] The SBU Captioned Photo Dataset (SBU) consists of over 1 million images with associated text descriptions. The SBU queries the Flickr using a huge number of pairs of query terms (objects, attributes, actions, stuff, and scenes). The querying method generates a huge number of noisy initial set of images with relevant text descriptions. Then, the images are filtered to ensure the high-relevance and visual-descriptive of images and textual descriptions. To encourage visual descriptiveness in the collection, SBU selects only those images with descriptions of satisfactory length based on observed lengths in visual descriptions.

A.2 Experimental Configurations

We provide details of training BALLAD and TACKLE in this section. The details of implementation on three benchmarks can be found in Table 1, 2, 3 for BALLAD, and Table 4, 5, 6 for TACKLE.

finetuning config	value
optimizer	SGD
visual learning rate	1e-5
linguistic learning rate	1e-5
weight decay	5e-4
momentum	0.9
batch size	512
epochs	50
sampler	instance-aware
learning rate schedule	cosine decay
adapting config	**value**
optimizer	SGD
learning rate	0.2
weight decay	5e-4
momentum	0.9
batch size	2048
epochs	10
sampler	class-aware
adapting ratio	$\lambda = 0.2$
learning rate schedule	cosine decay
Data augmentation	**value**
image size	224
random crop	scale=(0.5,1)
interpolation	BICUBIC
random horizontal flip	p=0.5

Table 2: Training configs of BALLAD on Places-LT.

finetuning config	value
optimizer	AdamW
visual learning rate	1e-5
linguistic learning rate	1e-6
weight decay	0.05
momentum	0.9
batch size	1024
epochs	400
sampler	instance-aware
learning rate schedule	cosine decay
adapting config	**value**
optimizer	AdamW
learning rate	0.2
weight decay	0.05
momentum	0.9
batch size	2048
epochs	20
sampler	class-aware
adapting ratio	$\lambda = 0.2$
learning rate schedule	cosine decay
Data augmentation	**value**
image size	224
random crop	scale=(0.5,1)
interpolation	BICUBIC
random horizontal flip	p=0.5

Table 3: Training configs of BALLAD on iNaturalist-2018.

training config	value
optimizer	AdamW
learning rate	1e-3
weight decay	0.05
momentum	0.9
batch size	1024
epochs	300
learning rate schedule	cosine decay
label smoothing	0.1
mixup	0.8
cutmix	1.0
drop path	0.1
repeated augmentation	True
Data augmentation	**value**
image size	224
color jitter	0.4
interpolation	BICUBIC
reprob	0.25
remode	pixel
recount	1.0

Table 4: Training configs of TACKLE on ImageNet-LT.

training config	value
optimizer	AdamW
learning rate	1e-3
weight decay	0.05
momentum	0.9
batch size	1024
epochs	300
learning rate schedule	cosine decay
label smoothing	0.1
mixup	0.8
cutmix	1.0
drop path	0.1
repeated augmentation	True
pretrained	ImageNet-1k
Data augmentation	**value**
image size	224
color jitter	0.4
interpolation	BICUBIC
reprob	0.25
remode	pixel
recount	1.0

Table 5: Training configs of TACKLE on Places-LT.

training config	value
optimizer	AdamW
learning rate	1e-3
weight decay	0.05
momentum	0.9
batch size	1024
epochs	300
learning rate schedule	cosine decay
label smoothing	0.1
mixup	0.8
cutmix	1.0
drop path	0.1
repeated augmentation	True
Data augmentation	value
image size	224
color jitter	0.4
interpolation	BICUBIC
reprob	0.25
remode	pixel
recount	1.0

Table 6: Training configs of TACKLE on iNaturalist-2018.

A.3 Algorithm of BALLAD

The overall algorithm of finetuning and balanced adapting of BALLAD framework is shown in Algorithm 1.

A.4 Text Prompting

Prompt engineering is initially proposed for knowledge probing in large pretrained language models [19, 24, 35, 40]. Prompting is adding extra instructions to task inputs to generate specific outputs from pretrained language model. In this paper, we utilize manually designed prompts following CLIP [36]. Specifically, a prompt template like *a photo of a* {CLASS} is adopted in experiments of BALLAD. As for TACKLE, according to the statement that text-prompts ensembling can improve model performance in CLIP [36], we leverage ensembled prompts in TACKLE. The emsembled prompts prefix consist of: *'a bad photo of a .', 'a photo of many .', 'a sculpture of a .', 'a photo of the hard to see .', 'a low resolution photo of the .', 'a rendering of a .', 'graffiti of a .', 'a bad photo of the .', 'a cropped photo of the .', 'a tattoo of a .', 'the embroidered .', 'a photo of a hard to see .', 'a bright photo of a .', 'a photo of a clean .', 'a photo of a dirty .', 'a dark photo of the .', 'a drawing of a .', 'a photo of my .', 'the plastic .', 'a photo of the cool .', 'a close-up photo of a .', 'a black and white photo of the .', 'a painting of the .', 'a painting of a .', 'a pixelated photo of the .', 'a sculpture of the .', 'a bright photo of the .', 'a cropped photo of a .', 'a plastic .', 'a photo of the dirty .', 'a jpeg corrupted photo of a .', 'a blurry photo of the .', 'a photo of the .', 'a good photo of the .', 'a rendering of the .', 'a in a video game.', 'a photo of one .', 'a doodle of a .', 'a close-up photo of the .', 'a photo of a .', 'the origami .', 'the in a video game.', 'a sketch of a .', 'a doodle of the .', 'a origami .', 'a low resolution photo of a .', 'the toy .', 'a rendition of the .', 'a photo of the clean .', 'a photo of a large .', 'a rendition of a .', 'a photo of a nice .', 'a photo of a*

Algorithm 1 Two-phases training of BALLAD

Require: Training samples $\{(\boldsymbol{I}, y)\}$, visual and language encoder $\mathcal{V}_{\text{enc}}$, $\mathcal{L}_{\text{enc}}$, linear adapter $\mathcal{LA}$

Initialize $\mathcal{V}_{\text{enc}}$, $\mathcal{L}_{\text{enc}}$ with web-data pretrained parameters Θ_v and Θ_l
for epoch $= 1, ..., N_A$ **do** ▷ Phase A
  for minibatch $B \in \{(\boldsymbol{I}, y)\}$ **do**
    $\boldsymbol{f}_v \leftarrow \mathcal{V}_{\text{enc}}(\boldsymbol{I}) \in \mathbb{R}^{d_v}$
    $\boldsymbol{T} \leftarrow tokenize(y)$
    $\boldsymbol{f}_l \leftarrow \mathcal{L}_{\text{enc}}(\boldsymbol{T}) \in \mathbb{R}^{d_l}$
    Project into embedding space $\mathbf{u}, \mathbf{v}$ as Eq.(1)
    Compute loss $\mathscr{L} \leftarrow \mathscr{L}_{v \to l} + \mathscr{L}_{l \to v}$ as Eq.(2)
    Update Θ_v and Θ_l
  end for
end for

Initialize Θ_{LA} randomly for $\mathcal{LA}$ ▷ Phase B
Freeze Θ_v and Θ_l
for epoch $= 1, ..., N_B$ **do**
  for minibatch $B \in \{Balanced(\boldsymbol{I}, y)\}$ **do**
    $\boldsymbol{f}_v \leftarrow \lambda \mathcal{LA}(\mathcal{V}_{\text{enc}}(\boldsymbol{I})) + (1-\lambda)\mathcal{V}_{\text{enc}}(\boldsymbol{I}) \in \mathbb{R}^{d_v}$
    $\boldsymbol{T} \leftarrow tokenize(y)$
    $\boldsymbol{f}_l \leftarrow \mathcal{L}_{\text{enc}}(\boldsymbol{T}) \in \mathbb{R}^{d_l}$
    Project into embedding space $\mathbf{u}, \mathbf{v}$ as Eq.(1)
    $p_i \leftarrow \frac{\exp(\mathbf{v}^\top \mathbf{u}_i)/\tau}{\sum_{j=1}^{K} \exp(\mathbf{v}^\top \mathbf{u}_j)/\tau}$
    Compute loss $\mathscr{L} \leftarrow CELoss(p, y)$
    Update Θ_{LA}
  end for
end for

weird .', 'a blurry photo of a .', 'a cartoon .', 'art of a .', 'a sketch of the .', 'a embroidered .', 'a pixelated photo of a .', 'itap of the .', 'a jpeg corrupted photo of the .', 'a good photo of a .', 'a plushie .', 'a photo of the nice .', 'a photo of the small .', 'a photo of the weird .', 'the cartoon .', 'art of the .', 'a drawing of the .', 'a photo of the large .', 'a black and white photo of a .', 'the plushie .', 'a dark photo of a .', 'itap of a .', 'graffiti of the .', 'a toy .', 'itap of my .', 'a photo of a cool .', 'a photo of a small .', 'a tattoo of the .',

B Related Work

Contrastive Vision-Language Model. Contrastive representation learning has been widely adopted to fulfill self-supervised pretraining in various AI domains[2, 3, 6, 14, 15, 32]. Recently, the intersection of vision and language [1, 7, 12, 23, 31, 39, 48] also experienced a revolution sparked by contrastive representation learning. Contrastive vision-language models like CLIP [36] and ALIGN [18] demonstrate promising zero-shot performances on various visual search and recognition tasks. Learning directly from natural language supervisions that contain rich visual concepts, they are very flexible and robust to distribution variations across different domains. The success of CLIP and ALIGN has enlightened many downstream vision-language tasks. For instance, DeCLIP [25] proposes to utilize self-, multi-view, and nearest-neighbor supervisions among the image-text pairs for data efficient pretraining of CLIP. On visual classification tasks, CLIP-Adapter [13] argues that fine-tuning contrastive vision-language models with linear adapters is a better alternative to prompt tuning. For video related tasks, VideoCLIP [46] performs contrastive pretraining with video-text pairs for zero-shot video-text understanding. ActionCLIP [43] presents a new "pretrain, prompt and fine-tune" paradigm leveraging pretrained vision-language models for zero-shot/few-shot action recognition. CLIP-It [30] designs a language-guided multimodal transformer based on CLIP to address query-focused video summarization. Moreover, CLIPort [41] combines CLIP with Transporter [49] to endow a robot with the ability of semantic understanding and spatial perception. In this paper, we demonstrate that contrastive vision-language models can also facilitate visual recognition under long-tailed class distribution setups if properly trained.

Long-Tailed Recognition. Long-tailed recognition [53] is a practical and challenging problem in vision domain. General visual models will suffer from severe performance degradation under such imbalanced class distributions. A great number of approaches [8, 9, 10, 17, 29, 34, 37, 44, 47, 51] have been proposed to address LTR from different perspectives. An intuitive solution is to directly re-balance the number of training samples across all classes [21, 54]. However, naively adjusting the skewness of training samples may lead to the overfitting of tail classes. Better alternatives include loss re-weighting [16, 20, 26] and logit adjustment [29, 50] based on label frequencies. Though efficacious for long-tailed distribution, above methods all sacrifice the performance of head classes at varying levels. To address the limitations, researchers turn to explore new network architectures and training paradigms. Typically, long-tail recognition models contain two key components – feature extractor and classifier. For each component, there are corresponding approaches by either designing better classifier [27, 42, 45] or learning reliable representations [28, 55]. In terms of new training frameworks, existing efforts seek to divide a one-stage training paradigm into two stages. For example, decoupled training approaches [21, 22] conduct representation learning and classifier training in a separate manner. Furthermore, ensemble learning schemes [52, 54] first learn multiple experts with different data sub-groups and then merge

Visual Backbone	ImageNet-LT		Places-LT		iNaturalist-2018	
	zero-shot	BALLAD	zero-shot	BALLAD	zero-shot	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)	-	-

Table 7: Top-1 accuracy of zero-shot CLIP and TACKLE-training.

their complementary knowledge to handle LTR. We borrow these ideas to optimize the fine-tuning of VL models.

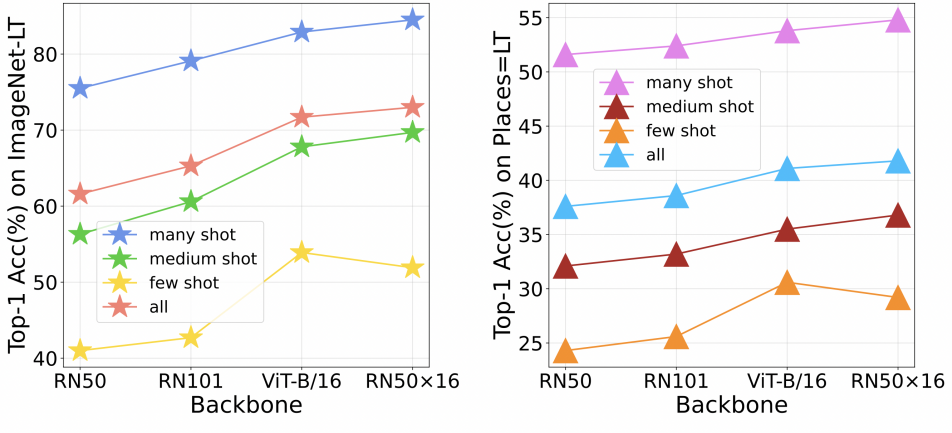
C More Ablations

C.1 Zero-shot Performance.

Comparison with zero-shot performance of CLIP in Table 7 shows promising improvements among all the backbones and all datasets, demonstrating the effectiveness of our BALLAD finetuning strategy.

C.2 Visual Backbones.

In BALLAD, we try CLIP with different visual backbones to explore its influence on final performance of TACKLE. We report the finetuning results of different backbones in Figure 1 on both ImageNet-LT and Places-LT benchmarks. When the visual backbone becomes deeper and larger, the finetuned performance is also gradually improved for *all*, *many-shot*, and *medium-shot* categories. Surprisingly, the Vision Transformer structure [▭] achieves the best accuracy in *few-shot* subset, probably owing to multi-head self-attention mechanism's ability in capturing minor features.



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

Figure 1: Comparisons between several visual backbones for ImageNet-LT (left) and Places-LT (right).

C.3 The Effectiveness of Pretrained Weights.

In Table 8, we validate the effectiveness of pretrained CLIP encoder weights in BALLAD compared with randomly initialized visual and linguistic encoders. All the four ablations are conducted on finetuning phase without data re-balancing for 50 epochs. The large gaps between random and pretrained CLIP initialization demonstrate the advantage of utilizing pretrained contrastive vision-language models. Besides, we find that visual encoder has much more influence than text encoder on the performance as random initialized vision encoder drops the accuracy close to zero. Note that poor performance of random initialization is primarily attributed to short training periods and pretrained vision-language weights fastening the convergence largely.

Vision	Language	many	medium	few	overall
random	*random*	0.3	0.0	0.0	0.1
random	*CLIP*	0.3	0.0	0.0	0.1
CLIP	*random*	36.8	2.9	0.0	15.6
CLIP	*CLIP*	**75.5**	**56.3**	**41.0**	**61.6**

Table 8: Ablations of pretrained vision-language weights on ImageNet-LT dataset. *CLIP* means using pre-trained weights as initialization and *random* represents random initialization.

C.4 Variants of Linear Adapter.

Since CLIP has dual encoders, the auxiliary linear adapter could be added to either or both of the two branches. As reported in Table 9, we try linear adapter for adapting visual and language encoders respectively. From the table, we can find that applying the linear adapter to the visual branch of CLIP achieves the best overall performance and is the optimal choice.

V-Adapter	*L-Adapter*	many	medium	few	overall
✓	-	**71.0**	**66.3**	**59.5**	**67.2**
-	✓	**71.0**	66.2	59.0	67.0
✓	✓	70.6	66.2	58.4	66.8

Table 9: Variants of linear adapter. *V-Adapter* and *L-Adapter* represents using linear adapter layer to adapt visual and language encoders respectively. All results are trained on ImageNet-LT for 10 epochs.

C.5 Where to balance.

Here, we compare re-balancing the long-tailed data distribution on either or both of two phases in BALLAD. The experiments are performed on ImageNet-LT and Places-LT datasets with ResNet-50-backboned CLIP. As mentioned earlier, *many-shot* categories dominate the feature space of long-tailed distribution. The performance drops of *many-shot* categories on both datasets, as reported in Table 10, suggest that balancing during Phase A tends to sacrifice *many-shot* representations. Since Phase A is mainly designed for updating representations on a new domain, we thereby abandon Phase-A data balancing.

When applying balancing strategies to Phase B alone, BALLAD can achieve a more balanced performance for different shots and improve the overall top-1 accuracy thanks to the rich features learned from Phase A.

Dataset	Balance Phase A	Balance Phase B	many	medium	few	overall
ImageNet-LT	-	-	**77.3**	57.4	39.0	62.6
	✓	-	76.6	58.4	42.7	63.3
	✓	✓	70.7	66.2	58.5	66.9
	-	✓	71.0	**66.3**	**59.5**	**67.2**
Places-LT	-	-	**52.7**	32.9	23.4	38.2
	✓	-	51.3	33.2	25.5	38.2
	✓	✓	44.6	46.7	44.1	45.5
	-	✓	46.7	**48.0**	**42.7**	**46.5**

Table 10: Ablations on where to employ balance strategies. On both ImageNet-LT and Places-LT, balance only in Phase B makes BALLAD to achieve the best performance.

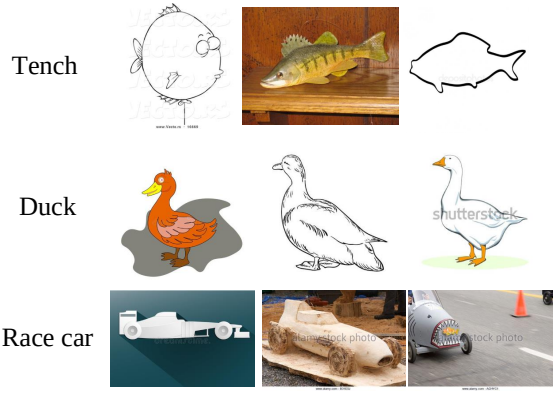


Figure 2: Web images retrieved by TACKLE using linguistic conceptual knowledge.

C.6 Conceptual Knowledge Transfer Visualization.

We visualize some web images retrieved by TACKLE as shown in Fig. 2. The conceptual language encoders of VL models provide linguistic hints to help conceive visual objects, which is the key of success in BALLAD. However, we show that TACKLE can transfer the conceptual knowledge without finetuning as the retrieved images complement target datasets from various perspectives (e.g., cartoon, mock-up, and design diagram as shown in Fig. 2). This proves that diverse images can facilitate visual encoders to understand real-world concepts more comprehensively.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, 2020.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[4] Craig Chambers, Ashish Raniwala, Frances Perry, Stephen Adams, Robert R Henry, Robert Bradshaw, and Nathan Weizenbaum. Flumejava: easy, efficient data-parallel pipelines. *ACM Sigplan Notices*, 45(6):363–375, 2010.

[5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019.

[8] Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. Reslt: Residual learning for long-tailed recognition. *arXiv preprint arXiv:2101.10633*, 2021.

[9] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 715–724, 2021.

[10] Qi Dong, Shaogang Gong, and Xiatian Zhu. Class rectification hard mining for imbalanced deep learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1851–1860, 2017.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[12] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6639–6648, 2019.

[13] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

[14] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[16] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6626–6636, 2021.

[17] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016.

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021.

[19] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

[20] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8420–8429, 2019.

[21] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1gRTCVFvB.

[22] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=OqtLIabPTit.

[23] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *arXiv preprint arXiv:1805.07932*, 2018.

[24] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[25] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.

[26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[27] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2970–2979, 2020.

[28] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019.

[29] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=37nvvqkCo5.

[30] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. In *Advances in Neural Information Processing Systems*, 2021.

[31] Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6087–6096, 2018.

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[33] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[34] Wanli Ouyang, Xiaogang Wang, Cong Zhang, and Xiaokang Yang. Factors in finetuning deep model for object detection with long-tail distribution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 864–873, 2016.

[35] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, 2019.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[37] Dvir Samuel and Gal Chechik. Distributional robustness loss for long-tail learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9495–9504, 2021.

[38] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[39] Lei Shi, Kai Shuang, Shijie Geng, Peng Su, Zhengkai Jiang, Peng Gao, Zuohui Fu, Gerard de Melo, and Sen Su. Contrastive visual-linguistic pretraining. *arXiv preprint arXiv:2007.13135*, 2020.

[40] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Eliciting knowledge from language models using automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, 2020.

[41] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Proceedings of the 5th Conference on Robot Learning (CoRL)*, 2021.

[42] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *Advances in Neural Information Processing Systems*, 2020.

[43] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

[44] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=D9I3drBz4UC.

[45] Tz-Ying Wu, Pedro Morgado, Pei Wang, Chih-Hui Ho, and Nuno Vasconcelos. Solving long-tailed recognition with deep realistic taxonomic classifier. In *European Conference on Computer Vision*, pages 171–189. Springer, 2020.

[46] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[47] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5704–5713, 2019.

[48] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019.

[49] Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, and Johnny Lee. Transporter networks: Rearranging the visual world for robotic manipulation. In *Proceedings of the 4th Conference on Robot Learning (CoRL)*, 2020.

[50] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2361–2370, 2021.

[51] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5409–5418, 2017.

[52] Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision. *arXiv preprint arXiv:2107.09249*, 2021.

[53] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*, 2021.

[54] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9719–9728, 2020.

[55] Linchao Zhu and Yi Yang. Inflated episodic memory with region self-attention for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4344–4353, 2020.