A Proofs

A.1 Proof of Lemma 2

Restated Lemma (Lemma 2). *For a balanced target distribution, where $P_t(y) = 1/K$ for all $y \in [K]$, we have:*

$$P_t(y|f_{\mathsf{ft}}(\mathbf{x}), f_{\mathsf{zs}}(\mathbf{x})) = \mathrm{softmax}(f_{\mathsf{ft}}(\mathbf{x}) + f_{\mathsf{zs}}(\mathbf{x}) - \pi_s - \pi_p)(y) \tag{13}$$

Proof. Denote the output $\mathbf{e} = f_{\mathsf{ft}}(\mathbf{x})$ and $\mathbf{z} = f_{\mathsf{zs}}(\mathbf{x})$. We first use the Bayes Rule to decompose $P_t(y|\mathbf{e}, \mathbf{z})$ into $P_t(\mathbf{e}, \mathbf{z}|y)$, $P_t(y)$ and $P_t(\mathbf{e}, \mathbf{z})$ in Eq. (14), then rewrite $P_t(\mathbf{e}, \mathbf{z}|y)$ in Eq. (15) according to Assumption 1. Focusing on label shift problem [33, 19, 28] where $P(\mathbf{x}|y)$ does not change, we derive Eq. (16)

$$\begin{aligned}
P_t(y|\mathbf{e}, \mathbf{z}) &= \frac{P_t(\mathbf{e}, \mathbf{z}|y)P_t(y)}{P_t(\mathbf{e}, \mathbf{z})} && (14)\\
&= P_t(\mathbf{e}|y)P_t(\mathbf{z}|y)\frac{P_t(y)}{P_t(\mathbf{e}, \mathbf{z})} && (15)\\
&= P_s(\mathbf{e}|y)P_p(\mathbf{z}|y)\frac{P_t(y)}{P_t(\mathbf{e}, \mathbf{z})} && (16)\\
&= \frac{P_s(y|\mathbf{e})P_s(\mathbf{e})}{P_s(y)}\frac{P_p(y|\mathbf{z})P_p(\mathbf{z})}{P_p(y)}\frac{P_t(y)}{P_t(\mathbf{e}, \mathbf{z})} && (17)\\
&= \frac{P_s(y|\mathbf{e})}{P_s(y)}\frac{P_p(y|\mathbf{z})}{P_p(y)}\frac{P_s(\mathbf{e})P_p(\mathbf{z})P_t(y)}{P_t(\mathbf{e}, \mathbf{z})} && (18)\\
& && (19)
\end{aligned}$$

Since $P_t(y) = 1/K$ is constant and $\mathbf{e}, \mathbf{z}$ are fixed, we can replace the terms that not rely on y with a constant C_1 in Eq. (20). Suppose the underlying class-probabilities $P_s(y|\mathbf{e}) \propto \exp(\mathbf{e}_y)$ and $P_p(y|\mathbf{z}) \propto \exp(\mathbf{z}_y)$ for $y \in [K]$. We replace $P_s(y) = \exp(\log P_s(y)) = \exp(\pi_s(y))$ and $P_p(y) = \exp(\log P_p(y)) = \exp(\pi_p(y))$ Denote the constant C_2 for normalizing P_s and P_p into probabilities , we get Eq. (21)

$$\begin{aligned}
P_t(y|\mathbf{e}, \mathbf{z}) &= \frac{P_s(y|\mathbf{e})}{P_s(y)}\frac{P_p(y|\mathbf{z})}{P_p(y)}C_1 && (20)\\
&= \exp(\mathbf{e} + \mathbf{z} - \pi_s - \pi_p)(y)\frac{C_1}{C_2} && (21)
\end{aligned}$$

Because the summation of $P_t(y|\mathbf{e}, \mathbf{z})$ is 1, $\frac{C_1}{C_2} = 1/\sum_{i\in[K]} \exp(\mathbf{e} + \mathbf{z} - \pi_s - \pi_p)(i)$. Therefore, we have:

$$\begin{aligned}
P_t(y|f_{\mathsf{ft}}(\mathbf{x}), f_{\mathsf{zs}}(\mathbf{x})) &= P_t(y|\mathbf{e}, \mathbf{z}) && (22)\\
&= \frac{\exp(\mathbf{e} + \mathbf{z} - \pi_s - \pi_p)_y}{\sum_{i\in[K]} \exp(\mathbf{e} + \mathbf{z} - \pi_s - \pi_p)_i} && (23)\\
&= \mathrm{softmax}(f_{\mathsf{ft}}(\mathbf{x}) + f_{\mathsf{zs}}(\mathbf{x}) - \pi_s - \pi_p)_y && (24)
\end{aligned}$$

□

A.2 Proof of Proposition 2

Restated Proposition (Proposition 2). *Suppose that the target distribution P_p is class-balanced. Let $h : \mathbb{R}^K \to \mathbb{R}^K$ be an arbitrary function that predicts labels using the outputs of the zero-shot model $f_{\mathsf{zs}}(\mathbf{x})$. Let the derived classifier be denoted as $f_h(\mathbf{x}) = h(f_{\mathsf{zs}}(\mathbf{x}))$. The classifier $f_{\mathsf{zs}} - \pi_p$ is better than any $f_h(\mathbf{x})$: $\mathcal{R}_t(f_{\mathsf{zs}} - \pi_p) \leq \mathcal{R}_t(f_h(\mathbf{x}))$.*

Dataset	Classes	Train size	Test size	Task
ImageNet	1,000	1.28M	50,000	Object-level
CIFAR100	100	50,000	10,000	Object-level
Caltech101	100	4,128	2,465	Object-level
DTD	47	2,820	1,692	Textures
EuroSAT	10	13,500	8,100	Satellite images
FGVCAircraft	100	3,334	3,333	Fine-grained aircraft
Flowers102	102	4,093	2,463	Fine-grained flowers
Food101	101	50,500	30,300	Fine-grained food
OxfordPets	37	2,944	3,669	Fine-grained pets
StanfordCars	196	6,509	8,041	Fine-grained car
SUN397	397	15,880	19,850	Scene-level
UCF101	101	7,639	3,783	Action
ImageNetV2	1,000	-	10,000	Robustness to collocation
ImageNet-Sketch	1000	-	50,889	Robustness to sketch domain
ImageNet-A	200	-	7,500	Robustness to adversarial attack
ImageNet-R	200	-	30,000	Robustness to multi-domains

Table 7: The detailed statistics of datasets for many-shot and few-shot learning.

Proof. Denote the output $\mathbf{z} = f_{\text{zs}}(\mathbf{x})$. Similar to Eq. (14)-Eq. (24), we have

$$\begin{aligned} P_t(y|\mathbf{z}) &= \frac{P_t(\mathbf{z}|y)P_t(y)}{P_t(\mathbf{z})} && (25)\\ &= \frac{P_p(\mathbf{z}|y)P_t(y)}{P_t(\mathbf{z})} && (26)\\ &= \frac{P_p(y|\mathbf{z})}{P_p(y)}\frac{P_t(y)}{P_t(\mathbf{z})} && (27)\\ &= \exp(\mathbf{z}-\pi_p)(y)/\sum_{i\in[K]}\exp((\mathbf{z}-\pi_p)(i)) && (28)\\ &= \text{softmax}(\mathbf{z}-\pi_p) = \text{softmax}(f_{\text{zs}}(\mathbf{x})-\pi_p) && (29) \end{aligned}$$

Therefore, we have:

$$\operatorname*{argmax}_{y\in\mathcal{Y}}(f_{\text{zs}}(\mathbf{x})-\pi_p)_y = \operatorname*{argmax}_{y\in\mathcal{Y}}\text{softmax}(f_{\text{zs}}(\mathbf{x})-\pi_p)_y = \operatorname*{argmax}_{y\in\mathcal{Y}} P_t(y|f_{\text{zs}}(\mathbf{x})) \tag{30}$$

Again, using Lemma 1, any other classifier $f_h(\mathbf{x})$ has higher risk than $f_{\text{zs}}(\mathbf{x})-\pi_p$, *i.e.*, $\mathcal{R}_t(f_{\text{zs}}-\pi_p) \leq \mathcal{R}_t(f_h(\mathbf{x}))$. □

B Experimental Details

B.1 Dataset details

Many-shot and few-shot datasets. For many-shot learning, we use ImageNet, CIFAR100, StanfordCars and SUN397 datasets. For few-shot learning, we evaluate models on 15 datasets. The details of each dataset are presented in Table 7.

Long-tail datasets. We use two standard long-tail benchmarks: Places365-LT and ImageNet-LT [29]. The skewness of a long-tailed training set is typically represented by imbalanced ratio, which is defined as $N_{\max}/N_{\min}$. $N_{\max}$ ($N_{\min}$) denotes the largest (smallest) number of instances per class. A larger imbalanced ratio means a more imbalanced training set. The test sets are divided into three splits: many-shot subset contains classes with > 100 images, medium-shot subset includes classes with ≥ 20 & ≤ 100 images, and few-shot subset covers classes with < 20 images. Details are listed in Table 8.

B.2 CLIP zero-shot

We use prompt ensembling of 80 prompts provided by CLIP [45] for ImageNet, CIFAR100, and Caltech101 to improve performance, *i.e.*, averaging the text embedding of many captions, *e.g.*., "a photo

Dataset	Size of all classes	Size of many classes	Size of medium classes	Size of few classes	Size of training samples	Imbalanced ratio
Places365-LT	365	131	163	71	62.5K	996
ImageNet-LT	1000	385	479	136	186K	256

Table 8: Details of long-tailed datasets.

of a $\{c_k\}$." and "an image of a $\{c_k\}$.". For OxfordPets, StanfordCars, Flowers102, Food101, FGVCAircraft, EuroSAT, UCF101, DTD and SUN397, we use the pre-defined prompt from CoOp [52].

B.3 Fine-tuned models

End-to-end and linear probe fine-tuning. We follow WiSE-FT [45] to implement fine-tuning. We initialize the classifier with the zero-shot classifier and the output of the image encoder Φ_v is normalized during fine-tuning. We fine-tune for a total of 10 epochs using AdamW [30] optimizer with default hyper-parameters $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$ and weight decay 0.1. We choose a batch size of 512. We use the same data augmentation and cosine-annealing learning rate schedule as [45].

B.4 Prompt tuning.

Prompt tuning like CoOp [52] automates prompt engineering by learning the prompt given few samples from downstream tasks. CoOp provides two options of prompt design: unified prompt that is shared among all classes and class-specific prompt that is different for each class. In this paper, we adopt the class-specific prompt design as the fine-tuned model to implement GLA . In specific, given the word embedding $\mathbf{t}_k^0$ initialized by zero-shot prompts, we aim to learn a collection of class-specific word embedding $\mathbf{R} = \{\mathbf{r}_k\}_{k=1}^K$, such that the text input $\mathbf{t}_k = \mathbf{t}_k^0 + \mathbf{r}_k$ minimizes the empirical risk: $\mathbf{R}^* = \mathrm{argmin}_\mathbf{R}\, \mathbb{E}_{\mathbf{x},y}[y \neq \mathrm{argmax}_i\, f(x;\mathbf{R})_i]$.

We adhere CoOp to use CLIP ResNet-50 as image encoder for few-shot classification. The word embedding $\mathbf{R}$ is initialized from zeros. For the m few-shot classification setting (where $m \in \{1, 2, 4, 8, 16\}$), we randomly sample m training and m validation points from the respective full datasets. For all few-shot datasets except ImageNet, the training epoch is set to 200 for 16/8 shots, 100 for 4/2 shots, and 50 for 1 shot. For ImageNet, the epoch is set to 50 for all shots. We fine-tune the prompt with SGD optimizer decayed by the cosine annrealing rule. The base initial learning rate and batch size are set to 10^{-4} and 32. When given an m-shot sample setting, we increase the learning rate and batch size by m times simultaneously to accelerate the training speed.

B.5 Estimation of the class prior

To estimate the log-probability of the pre-training distribution $\hat{\pi}_s = \log \mathbf{q}$, we utilize the optimization toolkit Cooper [13] from `https://github.com/cooper-org/cooper`. $\mathbf{q}$ is initialized as a uniformed distribution, $\mathbf{q}(y) = \frac{1}{K}$ for all $y \in [K]$. We use the standard SGD as the primal and dual optimizers for 2000 steps.

B.6 Long-tail learning baselines and training details

We compared with 5 long-tailed classification methods:

1. Standard ERM: We learn the model by standard empirical risk minimization on the long-tailed data.
2. Learnable Weight Scaling (LWS) [22]: We first learn the model by standard ERM, then fix the model and learn to re-scale the magnitude of the classifier using class-balanced sampling.
3. Logit Adjustment (LA) [33]: We first learn the model by standard ERM, then compensates the long-tailed distribution by subtracting a class-dependent offset to the model outputs.
4. Balanced Softmax (BS) [40] modifies the Softmax cross-entropy loss which explicitly accommodate the label distribution shift during optimization.

Model	Source	Target	
	ViT-B/32	ViT-B/16	ViT-L/14
Original zero-shot model $f_{\text{zs}}(\mathbf{x})$	63.4	68.8	75.6
Debiased zero-shot model $f_{\text{zs}}(\mathbf{x}) - \hat{\pi}_p$	**65.4**	**69.3**	**76.3**

Table 9: Estimated π_p is transferable across different backbones. $\hat{\pi}_p$ is estimated using CLIP ViT-B/32.

5. BALLAD [31] first fine-tunes the vision-language models via contrastive loss on long-tailed data, then freezes the backbone and finally employs an adapter to enhance the representations of tail classes with re-sampling strategies.

For all combinations of the fine-tuning baselines and long-tailed learning methods, visual backbones are initialized from CLIP-ResNet-50 and all classifiers are initialized by feeding prompt with class names to the text encoder. We use SGD for all experiments with a momentum of 0.9 for 50 epochs with batch size of 512. The initial learning rate is set to 1.6×10^{-3} which is decayed by the cosine annealing rule. To mitigate explosive gradients, we use the warmup learning rate equals to 10^{-5} during the first epoch. For the sake of fairness in comparison, all hyper-parameters of baselines are carefully searched using grid search on the validation set.

C Additional Experiments

C.1 Estimated label distribution is transferable

The estimated $\hat{\pi}_p$ should be transferable across different zero-shot models if they are trained on the same pre-training dataset. To confirm this, we estimate π_p using CLIP ViT-B/32 based zero-shot model, and use it to debias zero-shot models based on CLIP ViT-B/16 and ViT-L/14. Results are shown in Table 9, where our debiased zero-shot models based on CLIP ViT-B/16 and ViT-L/14 using $\hat{\pi}_p$ estimated from ViT-B/32 show clear performance gains over original zero-shot models.

C.2 Few-shot learning accuracy

We provide mean and standard deviation in Table 10 in for $\{1, 2, 4, 8, 16\}$ shots on all 11 few-shot learning datasets.

Dataset	1 shot	2 shots	4 shots	8 shots	16 shots
ImageNet	61.65 ± 0.15	62.64 ± 0.01	63.32 ± 0.07	64.51 ± 0.09	65.61 ± 0.03
Caltech101	89.08 ± 0.09	90.25 ± 0.25	90.98 ± 0.43	91.90 ± 0.21	92.58 ± 0.42
OxfordPets	87.79 ± 0.15	87.86 ± 0.21	88.22 ± 0.21	88.09 ± 0.27	89.53 ± 0.16
StanfordCars	60.00 ± 0.14	63.10 ± 0.42	66.25 ± 0.19	69.87 ± 0.09	73.95 ± 0.11
Flowers102	73.45 ± 0.60	81.00 ± 0.46	88.31 ± 0.65	92.89 ± 0.46	95.41 ± 0.32
Food101	78.41 ± 0.07	78.62 ± 0.07	78.68 ± 0.06	78.85 ± 0.19	79.54 ± 0.47
FGVCAircraft	20.22 ± 0.59	22.09 ± 0.37	24.65 ± 0.85	28.23 ± 0.44	31.99 ± 0.50
SUN397	64.29 ± 0.19	66.32 ± 0.16	68.01 ± 0.08	69.99 ± 0.18	71.64 ± 0.21
DTD	47.38 ± 1.23	50.75 ± 1.46	56.90 ± 0.20	62.73 ± 0.80	65.78 ± 0.49
EuroSAT	56.50 ± 1.34	67.26 ± 3.58	72.40 ± 2.43	77.59 ± 1.84	84.93 ± 1.89
UCF101	65.32 ± 0.17	68.42 ± 0.81	70.88 ± 0.50	74.23 ± 0.24	76.07 ± 0.03

Table 10: GLA Accuracy (%) with standard deviation of few-shot learning on 11 datasets.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022.

[2] Martin Arjovsky. *Out of distribution generalization in machine learning*. PhD thesis, New York University, 2020.

[3] Eric Bauer and Ron Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning*, 1999.

[4] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *Technical Report*, 2021.

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.

[7] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Prompt learning with optimal transport for vision-language models. *ICLR*, 2023.

[8] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[10] Thomas G Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems: First International Workshop*, 2000.

[11] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, 2004.

[12] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 1997.

[13] Jose Gallego-Posada and Juan Ramirez. Cooper: a toolkit for Lagrangian-based constrained optimization. `https://github.com/cooper-org/cooper`, 2022.

[14] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990.

[15] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

[16] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *CVPR*, 2021.

[17] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021.

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *NeurIPS Workshop*, 2014.

[19] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *CVPR*, 2021.

[20] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *UAI*, 2018.

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

[22] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020.

[23] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *CVPRW*, 2013.

[24] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[25] Ananya Kumar, Tengyu Ma, Percy Liang, and Aditi Raghunathan. Calibrated ensembles can mitigate accuracy tradeoffs under distribution shift. In *UAI*. PMLR, 2022.

[26] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *NeurIPS*, 2017.

[27] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *NeurIPS*, 2020.

[28] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *ICML*, 2018.

[29] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019.

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[31] Teli Ma, Shijie Geng, Mengmeng Wang, Jing Shao, Jiasen Lu, Hongsheng Li, Peng Gao, and Yu Qiao. A simple long-tailed recognition baseline via vision-language model. *arXiv preprint arXiv:2111.14745*, 2021.

[32] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[33] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021.

[34] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics & Image Processing*, 2008.

[35] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012.

[36] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*, 2021.

[37] 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 *ICML*, 2021.

[38] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019.

[39] William J Reed. The pareto, zipf and other power laws. *Economics letters*, 2001.

[40] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *NeurIPS*, 2020.

[41] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[42] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *NeurIPS*, 2020.

[43] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *NeurIPS*, 2020.

[44] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 2019.

[45] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *CVPR*, 2022. https://arxiv.org/abs/2109.01903.

[46] Tong Wu, Ziwei Liu, Qingqiu Huang, Yu Wang, and Dahua Lin. Adversarial robustness under long-tailed distribution. *CVPR*, 2021.

[47] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[48] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *TMLR*, 2022.

[49] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *ECCV*, 2022.

[50] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *CVPR*, 2021.

[51] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 2004.

[52] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022.

[53] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. *arXiv preprint arXiv:2205.14865*, 2022.

[54] Beier Zhu, Yulei Niu, Saeil Lee, Minhoe Hur, and Hanwang Zhang. Debiased fine-tuning for vision-language models by prompt regularization. *AAAI*, 2023.