Cross-Domain Empirical Risk Minimization for Unbiased Long-Tailed Classification

Beier Zhu[1], Yulei Niu[1]*, Xian-Sheng Hua[2], Hanwang Zhang[1]

[1]Nanyang Technological University [2]Damo Academy, Alibaba Group
beier002@e.ntu.edu.sg, yn.yuleiniu@gmail.com, xiansheng.hxs@alibaba-inc.com, hanwangzhang@ntu.edu.sg

Abstract

We address the overlooked unbiasedness in existing long-tailed classification methods: we find that their overall improvement is mostly attributed to the biased preference of "tail" over "head", as the test distribution is assumed to be balanced; however, when the test is as imbalanced as the long-tailed training data—let the test respect Zipf's law of nature—the "tail" bias is no longer beneficial overall because it hurts the "head" majorities. In this paper, we propose Cross-Domain Empirical Risk Minimization (xERM) for training an unbiased model to achieve strong performances on both test distributions, which empirically demonstrates that xERM fundamentally improves the classification by learning better feature representation rather than the "head vs. tail" game. Based on causality, we further theoretically explain why xERM achieves unbiasedness: the bias caused by the domain selection is removed by adjusting the empirical risks on the imbalanced domain and the balanced but unseen domain.

Introduction

When the training data distribution is long-tailed, *e.g.*, thousands of samples per "head" class vs. just a few samples per "tail" class, the resultant classification model will inevitably downplay the minor "tail" and overplay the major "head". Therefore, to lift up the "tail" performance, recent methods focus on the trade-off between the "head" and "tail" contributions, *e.g.*, sample re-weighting (Cui et al. 2019; Lin et al. 2017), balanced training loss (Kang et al. 2020; Zhang et al. 2021b), and "head" causal effect removal (Tang, Huang, and Zhang 2020). As you may expect, the "tail" rise costs the "head" fall; but fortunately, the overall gain is still positive. The reasons are two-fold. First, thanks to the abundant "head" training samples, the strong "head" features can afford the trade-off, making the "head" performance drop much less significant than the "tail" rise. Second, the test distribution is often assumed to be balanced, *i.e.*, "head" and "tail" have the same number of samples, so the "head" fall is not amplified by the balanced sample number.

We conjecture that the above "overall improvement" on the balanced test may cover up the fact that today's long-tailed classification methods are still *biased* and far from truly "improved". Figure 1 (a.1) and (b.1) show an evidence. Regardless of the test distributions, balanced or imbalanced, an imbalanced model (vanilla cross-entropy trained, XE, orange line) always predicts the class distribution shaped as its own model prior, which is long-tailed. As a comparison, a balanced model (a SOTA method TDE (Tang, Huang, and Zhang 2020), blue line) always predicts the distribution to cater the tail. This observation implies that both of them are biased: the imbalanced model is "head" biased; the balanced one is "tail" biased, which merely replaces the "head" bias with the "tail" bias. Therefore, we expect an "unbiased" model's prediction to treat each class fairly and respects both test distributions, *i.e.*, the prediction distribution follows the shape of the ground-truth one (*e.g.*, green line). Figure 1 (a.2) and (b.2) provide another evidence based on the true positive (TP) and the false positive (FP) statistics. We observe that the balanced model TDE tends to predict more samples as tail classes. On the contrary, the imbalanced model XE prefers to predict more samples as head classes. The preference on head or tail classes comes with a large number of FP samples, which lowers the precision score. In contrast, our xERM has no significant preference between head and tail classes, which achieves high precision under both test distributions. Only through the above unbiasedness, we can diagnose whether a model truly improves long-tailed classification, but not by playing a "prior flipping" game.

In fact, it is challenging to achieve unbiasedness across the two completely different test distributions due to the domain selection issue (Pearl and Bareinboim 2011): the two distributions would never co-exist in training — once one is observed, the other is gone — let alone adjusting a fair prior for both of them. In this paper, we propose a novel training paradigm called Cross-Domain Empirical Risk Minimization (xERM) to tackle this challenge. xERM consists of two ERM terms, which are the cross-entropy losses with different supervisions: one is the ground-truth label from the seen imbalanced domain, and the other is the prediction of a balanced model. The imbalanced domain empirical risk encourages the model to learn from the ground-truth annotations from the imbalanced distribution, which favors the head classes. The balanced domain empirical risk encourages the model to learn from the prediction of a balanced model, which imagines a balanced distribution and prefers the tail classes. These two risks are weighted to take advantage of

*Corresponding author
Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

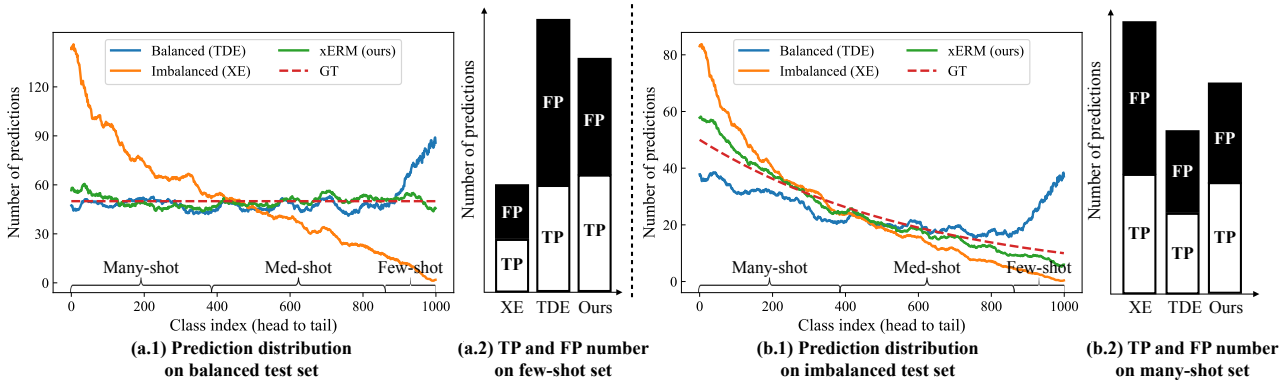


Figure 1: Prediction distributions of imbalanced model (XE), balanced model (TDE) and our xERM model on (a.1) balanced test and (b.1) imbalanced test (long-tailed) of ImageNet-LT. Interestingly, the abrupt tail rise of the blue lines shows an obvious "tail" preference. The true positive (TP) and false positive (FP) number of the predictions on (a.2) few-shot subset of the balanced test set and (b.2) many-shot subset of the imbalanced test set.

both, *i.e.*, protect the model training from being neither too "head biased" nor too "tail biased", and can achieve the unbiasedness (*cf.* Figure 1). Furthermore, we propose a causal theory to ground xERM, which removes the confounding bias caused by domain selection.

We conduct experiments on several long-tail classification benchmark datasets: CIFAR100-LT (Cui et al. 2019), Places365-LT (Liu et al. 2019), ImageNet-LT (Liu et al. 2019), and iNaturalist 2018 (Van Horn et al. 2018). Experimental results show that xERM outperforms previous state-of-the-arts on both long-tailed and balanced test sets, which demonstrates that the performance gain is not from catering to the tail. Further qualitative studies show that the xERM helps with better feature representation.

We summarize our contributions as follows:

- We point out the overlooked *unbiasedness* in long-tailed classification: models should perform well on both imbalanced and balanced test distributions. Compared to the conventional evaluations on the balanced test only, the unbiasedness thoroughly diagnoses whether a model is spuriously improved by simply flipping the bias.
- We propose xERM that achieves the unbiasedness. On various long-tailed visual recognition benchmarks, we show that xERM constantly achieves strong performances across different test distributions.
- We provide a theoretical analysis for xERM which reduces the confounding bias based on the causal theory.

Related Work

Long-tailed classification. Early works focus on re-balancing the contribution of each class in the training phase, which can be divided into two strategies: re-sampling the data to balance the class frequency (Buda, Maki, and Mazurowski 2018; Drumnond 2003; Shen, Lin, and Huang 2016; Sarafianos, Xu, and Kakadiaris 2018; Japkowicz and Stephen 2002), and re-weighting the loss of classes (Tan et al. 2020; Cui et al. 2019; Lin et al. 2017; Khan et al. 2017) to increase the importance of tail classes. Nevertheless, both strategies suffer from under-fitting/over-fitting problem to head/tail classes. Recent studies shows a trend of decoupling long-tailed classification into two separate stages: representation learning and classifier learning. Decouple (Kang et al. 2020) learns high-quality representation with natural (instance-balanced) sampling, and achieve strong classification performance by adjusting classifier with balanced learning. Zhou et al. (2020) arrives at a similar conclusion by proposing a bilateral-branch network, where one branch is trained with natural sampling and the other uses balanced sampling. The decoupled learning idea is also adopted in (Menon et al. 2021; Hong et al. 2021; Tang, Huang, and Zhang 2020; Zhang et al. 2021b), where different classifier adjustment methods are proposed. There is a line of post-hoc logits adjustment by subtracting the training prior from the predicted logits (Menon et al. 2021; Hong et al. 2021). Tang, Huang, and Zhang (2020) removes the "head" causal effect to get the adjusted logits. DisAlign (Zhang et al. 2021b) modifies the original logits by adding an extra learnable layer to the output layer. Apart from decoupled learning, Liu et al. (2020) addresses the long-tailed problem by transferring head distributions to tail; distilling balanced student models from imbalanced trained teacher models (He, Wu, and Wei 2021; Zhang et al. 2021a). Although all the above methods manifest an overall accuracy improvement, they are indeed biased.

Causal inference. Causal inference (Pearl, Glymour, and Jewell 2016; Peters, Janzing, and Schölkopf 2017) has been studied in a wide spectrum, such as economics, politics and epidemiology (Heckman 1979; Keele 2015; Robins 2001). Recently, causal inference has also shown promising results in various computer vision tasks, *e.g.*, counterfactual training or inference in VQA (Niu et al. 2021; Chen et al. 2020, 2021), back-door and front-door adjustment for vision-language tasks (Yang, Zhang, and Cai 2021; Qi et al. 2020; Yang et al. 2021), and counterfactual inference for zero-shot recognition (Yue et al. 2021). Our algorithm can be viewed as a practical implementation and complement of the recent empirical risk minimization of learning causal effect (Jung, Tian, and Bareinboim 2020) for confounding effect removal.

Bias mitigation. A bias mitigation work related to us is LfF (Nam et al. 2020). LfF used a reweighting-based method to deal with bias in out-of-distribution robustness, and used the cross-entropy loss to calculate the weights. The main differences are as follows. First, LfF focused on biased dataset with spurious correlations, while we focused on long-tailed task with label prior shift. Second, LfF weights the losses at the sample-level, while xERM weights the losses at the supervision-level. Third, LfF calculates the weights using a biased model and a target debiased model, while xERM calculates the weights using fixed balanced and imbalanced models which can be merged into one. Another related work (Niu and Zhang 2021) also focused on achieving a good trade-off between in-distribution and out-of-distribution samples for robust QA. In this paper, we generalize it to long-tailed classification and further provided a causal perspective and justification of the unbiased learning framework.

The Algorithm: xERM

Input: Long-tailed training samples denoted as the the pairs $\{(x, y)\}$ of an image x and its one-hot label y.
Output: Unbiased classification model f.

Step 1: Learn an imbalanced and a balanced model. We learn a conventional model $p^{\text{imba}}(y|x)$ on the imbalanced training data as the imbalanced model, which favors the "head" bias. Then, we learn a balanced model $p^{\text{ba}}(y|x)$, which favors the "tail" bias. The implementations of the two models are open. See Experiments and Appendix for our choices.

Step 2: Estimate the adjustment weights.

Imbalanced Domain ER weight:

$$w^{\text{imba}} = \frac{(XE^{\text{imba}})^{\gamma}}{(XE^{\text{imba}})^{\gamma} + (XE^{\text{ba}})^{\gamma}}, \tag{1}$$

Balanced Domain ER weight:

$$w^{\text{ba}} = 1 - w^{\text{imba}} = \frac{(XE^{\text{ba}})^{\gamma}}{(XE^{\text{imba}})^{\gamma} + (XE^{\text{ba}})^{\gamma}}, \tag{2}$$

where $XE^{\text{imba}} = -\sum_i y_i \log p^{\text{imba}}(y_i|x)$ is the cross-entropy of the imbalanced model's predictions (similar for XE^{ba}), the subscript i denotes the i-th class label, and $\gamma > 0$ is a scaling hyper-parameter.

Step 3: Minimize the cross-domain empirical risk

Imbalanced Domain ER: $$\mathcal{R}^{\text{imba}}(f) = -w^{\text{imba}} \sum_i y_i \log f_i(x), \tag{3}$$

where y_i and f_i are the ground-truth and the predicted label for i-th class, respectively.

Balanced Domain ER: $$\mathcal{R}^{\text{ba}}(f) = -w^{\text{ba}} \sum_i \hat{y}_i \log f_i(x), \tag{4}$$

where $\hat{y}_i = p^{\text{ba}}(y_i|x)$ denotes the balanced prediction for i-th class. The overall empirical risk minimization:

$$\mathcal{R}(f) = \mathcal{R}^{\text{imba}}(f) + \mathcal{R}^{\text{ba}}(f). \tag{5}$$

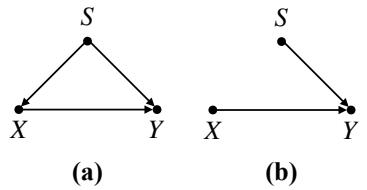


Figure 2: (a) The causal graph for long-tailed classification. The spurious correlation is induced as a backdoor $X \leftarrow S \rightarrow Y$ via the selection variable S. (b) The post-interventional model cuts off the arrow towards X by $do(X)$.

Justification

Causal Graph. We first construct the causal graph for long-tailed classification to analyze the causal relations between variables. As shown in Figure 2 (a), the causal graph contains three nodes: X: image, Y: prediction, S: selection variable. Each value of S corresponds to a domain, and switching between the domains will be represented by conditioning on different values of the S variable. In long-tailed classification, let $S = 0$ denote the balanced domain and $S = 1$ denote the imbalanced (long-tailed) domain.[1]

$X \rightarrow Y$. This path indicates that the model predicts the label based on the image content.

$S \rightarrow X$. This path indicates that an image is sampled according to the selected domain, *e.g.*, imbalanced domain is prone to include images from head classes.

$S \rightarrow Y$. This path indicates that the predicted label distributions follow their own training domain prior. For example, in Figure 1, an imbalanced model XE always predicts a long-tailed label distribution regardless of the test distributions, and the balanced model TDE shows an obvious "tail" preference across different test distributions.

Note that the back-door path $X \leftarrow S \rightarrow Y$ contributes a spurious correlation between X and Y (Pearl, Glymour, and Jewell 2016), where S acts as a confounder. The existence of back-door path is the key challenge for achieving unbiasedness across balanced and imbalanced test data. Simply learning by $P(Y|X = x)$ will inevitably include such spurious correlation, which is not what we seek to estimate, and lead to a biased model. Therefore, we instead estimate $P(Y|do(X = x))$ which can eliminate the confounding bias to obtain an unbiased model.

Empirical Risk on Intervened Distribution. The empirical risk minimization for learning causal effect $P(Y|do(X = x))$ has been well formulated in (Jung, Tian, and Bareinboim 2020). We define the empirical risk $\mathcal{R}$ of the estimator f on the intervened distributions as:

$$\begin{aligned}\mathcal{R}(f) &= \mathbb{E}_{x \sim P(X), y \sim P(Y|do(X=x))} \mathcal{L}(y, f(x)) \\ &= \sum_x \sum_y \mathcal{L}(y, f(x)) P(y|do(x)) P(x),\end{aligned} \tag{6}$$

[1]Technically, the possible values of S are infinite and far more than only confined to 0 (balance) or 1 (long-tail). However, a fine-grained definition of S leads to heavy computation, which is impractical for training. In this paper, we focus on the long-tailed and balanced domains.

where $\mathcal{L}(y, f(x))$ is formulated as the standard cross-entropy. By do-operation (causal intervention), we cut off the in-paths to X and transform the causal graph in Figure 2 (a) to Figure 2 (b). Different from traditional empirical risk that samples (x, y) from $P(X, Y)$, we sample (x, y) from the intervened distribution $P(do(X), Y)$. According to the rules of do-calculus (Pearl 2000), we have $P(do(X = x)) = P(X = x)$. Therefore, we can implement the sampling process drawn from the intervened distribution as two steps. First, we sample x from $P(X)$. Second, we sample y from the intervened conditional probability $P(Y|do(X = x))$. In short, we denote $P(X=x)$ as $P(x)$ and $P(Y=y|do(X=x))$ as $P(y|do(x))$.

The backdoor adjustment (Pearl 2000; Pearl, Glymour, and Jewell 2016) allows us to remove $do(x)$ in $P(y|do(x))$ for easier estimation. We rewrite $P(y|do(x))$ *w.r.t.* the confounder S as:

$$\begin{aligned} P(y|do(x)) &= \sum_{S=s\in\{0,1\}} P(y|x, S = s)P(S = s) \\ &= \frac{P(x, y, S = 1)}{P(x|S = 1)} + \frac{P(x, y, S = 0)}{P(x|S = 0)}. \end{aligned} \tag{7}$$

A more detailed explanation of Eq. (7) is provided in Appendix. Here $S = 1$ denotes the seen imbalanced domain, *i.e.*, training samples, and $S = 0$ denotes the unobserved balanced domain. Combining Eq. (7) and Eq. (6), we have:

$$\begin{aligned} \mathcal{R}(f) &= \sum_{(x,y)} \sum_{s\in\{0,1\}} \mathcal{L}(y_s, f(x)) \frac{P(x)}{P(x|S = s)} P(x, y, s) \\ &= \frac{1}{N} \sum_{(x,y)} [\underbrace{\mathcal{L}(y_{s=1}, f(x)) \frac{P(x)}{P(x|S = 1)}}_{\text{Imbalanced Domain ER}} \\ &\quad + \underbrace{\mathcal{L}(y_{s=0}, f(x)) \frac{P(x)}{P(x|S = 0)}}_{\text{Balanced Domain ER}}], \end{aligned} \tag{8}$$

where $y_{s=1}$ denotes the ground-truth label from the seen imbalanced domain and $y_{s=0}$ denotes the supervision from the unseen balanced domain. Here remain two questions. First, how to obtain $y_{s=0}$. As mentioned above, we cannot directly draw samples $(x, y_{s=0})$ from the unseen domain $S=0$. Fortunately, thanks to the causality-based framework (Tang, Huang, and Zhang 2020), this can be simulated by *counterfactual inference* that deducts the "head" prior effect from the factual effect. In this way, we can obtain the counterfactual $(x, y_{s=0})$ based on the factual $(x, y_{s=1})$. In particular, $y_{s=0}$ is estimated from the balanced classifier $p^{\text{ba}}(y|x)$ described in Step 1 and denoted as $\hat{y}$ in Step 3 of the algorithm. Second, how to obtain the sample weights $\frac{P(x)}{P(x|S)}$. Since $\frac{P(x)}{P(x|s)} = \frac{P(s)}{P(s|x)}$ and we assume $P(S = 1) = P(S = 0)$. Therefore, we have $\frac{P(x)}{P(x|s)} \propto \frac{1}{P(s|x)}$. In essence, they are *inverse* probabilities of how likely the sample x is from the imbalanced training domain ($S = 1$) and from the balanced domain ($S = 0$). Inspired by this, we propose to use the cross-entropy loss XE to delineate the discrepancy between the imbalanced and balanced domains and estimate the weights:

$$\begin{aligned} w^{\text{imba}} &\propto \frac{1}{P(S = 1|x)} \propto (XE^{\text{imba}})^{\gamma}, \\ w^{\text{ba}} &\propto \frac{1}{P(S = 0|x)} \propto (XE^{\text{ba}})^{\gamma}. \end{aligned} \tag{9}$$

Intuitively, the above equation says that if x does not fit to the imbalanced model $p^{\text{imba}}(y|x)$ (large XE^{imba}) , it is not likely from the imbalanced training domain (small $P(S = 1|x)$), and vice versa. By normalizing the weights to sum 1, we can easily derive Eq. (1) and (2) from Eq. (9). A large hyper-parameter $\gamma > 0$ would push w towards 0 or 1 and we studied the effect of γ in ablation studies.

Experiments

Setup

Datasets. We conducted experiments on four long-tailed classification datasets: CIFAR100-LT (Cui et al. 2019), Places365-LT (Liu et al. 2019), ImageNet-LT (Liu et al. 2019), and iNaturalist 2018 (Van Horn et al. 2018). The imbalance ratio (IB) is defined as $N_{\max}/N_{\min}$, where $N_{\max}$ ($N_{\min}$) denotes the largest (smallest) number of class samples. A larger IB denotes a more imbalanced split. See Appendix for further details.

Backbones. For fair comparisons, we used ResNet-32 (He et al. 2016) as the backbone for CIFAR100-LT, ResNet-{50, 152} for Places365-LT, ResNeXt-50 for ImageNet-LT and ResNet-50 for iNaturalist18. For Places365-LT, we pretrained the backbones on ImageNet dataset as in Liu et al. (2019).

Comparison methods. We compare the proposed xERM models with state-of-the-arts such as BBN (Zhou et al. 2020), LDAM (Cao et al. 2019), LFME (Xiang, Ding, and Han 2020), BKD (Zhang et al. 2021a), cRT, τ-Norm, LWS (Kang et al. 2020), LADE, PC (Hong et al. 2021), LA (Menon et al. 2021),TDE (Tang, Huang, and Zhang 2020), DiVE (He, Wu, and Wei 2021) and DisAlign (Zhang et al. 2021b).

The above mentioned methods use single models. Note that there is also a line of using ensemble models for long-tailed classification, *e.g.*, RIDE (Wang et al. 2020) and CBD (Iscen et al. 2021). For fair comparisons, we will not include their results in the experiments. But, in Appendix, we show that our xERM framework can be easily applied to RIDE for validating its effectiveness on ensemble models.

Implementation. We chose TDE and PC as baselines to implement xERM for two reasons. First, xERM requires an imbalanced model and a balanced model. Since TDE and PC are post-hoc logits adjustment methods, we can also obtain the imbalanced model effortlessly by disabling the post-hoc operation, which reduces the computation cost. Otherwise, we have to separately pre-train two models for the purposes. The second is performance. Balanced model with higher performance leads to a better xERM model. TDE and PC achieve stronger performance comparing to previous state-of-the-arts. TDE first projects the feature from the penultimate layer to the orthogonal direction of the mean feature, then predicts the logits based on the projected feature. PC adjusts logits by subtracting the training prior from the predicted logits. We set the scaling parameter γ in Eq. (9) to 2 for

Method	Acc	Recall			Precision			F1		
		Many	Med	Few	Many	Med	Few	Many	Med	Few
XE	40.5	68.5	39.9	8.6	38.5	50.9	**52.2**	49.3	44.7	14.8
BBN	42.3	61.5	41.7	20.2	46.6	49.7	35.1	53.0	45.3	25.6
cRT	44.4	62.9	45.1	21.8	46.9	48.2	36.7	53.7	46.6	27.4
LADE	45.0	60.1	43.3	**29.3**	**52.2**	50.9	32.0	55.8	46.8	30.5
TDE	44.1	63.9	46.9	17.8	46.6	45.5	34.1	53.9	46.2	23.4
PC	45.5	62.2	45.6	25.8	50.6	48.5	32.5	55.8	47.0	28.7
xERM$_{TDE}$	46.2	**69.0**	**48.8**	16.7	46.4	49.3	45.5	55.5	**49.0**	24.4
xERM$_{PC}$	**46.9**	65.5	45.4	26.8	51.1	**51.5**	35.9	**57.4**	48.3	**30.7**

Table 1: Comparisons on CIFAR100-LT-IB-100.

Method	Acc	Recall			Precision			F1		
		Many	Med	Few	Many	Med	Few	Many	Med	Few
XE	49.0	68.6	42.9	15.0	46.9	**59.1**	**60.7**	55.7	49.7	24.1
τ-Norm	49.6	61.8	46.2	27.4	52.2	48.5	43.7	56.6	47.3	33.7
LWS	49.9	60.2	47.2	30.3	53.0	49.1	41.3	56.4	48.1	35.0
LADE	51.7	62.6	49.0	30.4	55.3	50.5	41.2	58.7	49.7	34.9
DiVE	53.1	64.1	50.4	**31.5**	-	-	-	-	-	-
DisAlign	53.4	61.3	**52.2**	31.4	-	-	-	-	-	-
PC	48.9	60.4	46.7	23.8	56.3	49.7	32.0	58.3	48.2	27.3
TDE	51.8	62.7	49.0	31.4	**57.3**	52.3	39.5	59.9	50.6	35.0
xERM$_{PC}$	53.2	67.6	49.8	24.0	53.1	55.0	52.4	59.3	51.9	33.0
xERM$_{TDE}$	**54.1**	**68.6**	50.0	27.5	53.5	57.3	52.0	**60.1**	**53.4**	**36.0**

Table 2: Comparisons on ImageNet-LT.

CIFAR100-LT, 5 for Places365-LT and 1.5 for other datasets. See Appendix for more details.

Evaluations on Balanced Test Set

Settings. We evaluated the proposed xERM models on balanced test set (*i.e.* IB=1). All networks were trained for 200 epochs on CIFAR100-LT, 30 epochs on Places365-LT, and 90 epochs on ImageNet-LT and iNaturalist18.

Metrics. We used the top-1 accuracy as an overall metric for all the datasets. To further evaluate the unbiasedness, we divided the test set of CIFAR100-LT-IB-100 and ImageNet-LT into three subsets according to the number of samples in each class: many-shot (categories with $>$100 images), medium-shot (categories with 20$\sim$100 images), and few-shot (categories with $<$20 images). Previous works focused on the macro-average recall for evaluation, especially on the few-shot subset. However, recall alone is not enough to measure the performance on each subset. A trivial solution to achieve a high recall on the few-shot subset is predicting all the samples to the few-shot classes. In this case, the performance on few-shot subset is not reliable. Therefore, we further reported the macro-average precision and the F1 scores for a more comprehensive evaluation. The results are summarized in Table 1-3. Further evaluations on Places365-LT and iNaturalist18 are in Appendix.

Results on top-1 accuracy. Overall, xERM outperforms state-of-the-art methods in terms of the top-1 accuracy on all datasets. In particular, compared to the balanced models, xERM model consistently outperforms TDE by 2.1%, 2.3%, 1.1%, and 1.4%, and PC by 1.4%, 4.3%, 2.2% and 1.1% on CIRAR100-LT-IB-100, ImageNet-LT, Places365-LT and iNaturalist18. These results demonstrate the effectiveness and generalizability of our xERM.

CIFAR100LT			Places365-LT			iNaturalist18	
Method	IB		Method	Backbone		Method	Acc
	50	10		R50	R152		
XE	46.5	58.9	XE	28.0	30.5	XE	59.3
LDAM	46.6	58.7	LFME	-	36.2	τ-Norm	65.6
BBN	47.0	59.1	LWS	36.2	37.6	LWS	65.9
cRT	48.7	59.8	τ-Norm	36.6	37.9	BBN	66.3
BKD	49.6	61.3	BKD	-	38.4	LA	66.4
LADE	50.5	61.7	LADE	36.5	38.8	LDAE	66.7
DiVE	51.1	62.0	DisAlign	37.8	**39.3**	BKD	66.8
TDE	50.3	59.6	TDE	37.2	38.1	TDE	63.1
PC	49.7	60.9	PC	36.1	38.3	PC	66.2
xERM$_{TDE}$	**52.8**	62.2	**xERM**$_{TDE}$	**38.3**	39.0	**xERM**$_{TDE}$	64.5
xERM$_{PC}$	51.1	**62.5**	**xERM**$_{PC}$	**38.3**	**39.3**	**xERM**$_{PC}$	**67.3**

Table 3: The top-1 accuracy on CIFAR100-LT-IB-{50,10}, Places365-LT, and iNaturalist18.

Results on recall. When it comes to the recall performances on the subsets, *all of the previous methods enhance the overall accuracy at the expense of a large head performance drop.* For example, in Table 1, although LADE (Hong et al. 2021) achieves a high top-1 accuracy of 45.0% and few-short recall of 29.3%, its many-shot recall encounters with a significant decrease of 8.4% compared to the baseline XE. In contrast, our xERM achieves the highest overall performance with a competitive head performance against XE. Surprisingly, xERM$_{TDE}$ is the first to beat XE w.r.t the recall performances on both CIFAR100LT-IB-100 and ImageNet-LT. These results demonstrate that our xERM achieves the unbiaseness from the factual and the counterfactual model, *i.e.*, improving the overall accuracy without scarifying the head recall.

Results on precision and F1 score. We further reported precision and F1 score to give a more comprehensive evaluation. According to the results on CIFAR100-LT-IB-100 and ImageNet-LT, the *low precision and high recall* are observed on (1) *many*-shot subset for the imbalanced model (XE); (2) *few*-shot set for the balanced models. This indicates that the imbalanced models are "head" biased and the balanced models are "tail" biased. In contrast, on ImageNet-LT, our xERM$_{PC}$ achieves both high recall and precision (recall: 67.6%, precision: 53.1%) on the many-shot subset compared to factual model XE that achieves high recall with lower precision (recall: 68.6%, precision: 46.9%), and achieves high recall and precision on the few-shot subset (recall: 24.0%, precision: 52.4%) compared to PC (Hong et al. 2021) that achieves competitive recall with lower precision (recall: 23.8%, precision: 32.0%). These results demonstrate the unbiasedness of our proposed method. Besides, an unbiased model should have a balanced precision over the three subsets. For example, in Table 2, our xERM$_{PC}$ manifests 53.1%, 55.0% and 52.4% over many-shot, medium-shot and few-shot precision with a small variance, while XE manifests 46.9%, 59.1% and 60.7% and PC manifests 56.3%, 49.7% and 32.0% with large variances. As a result, our xERM achieves the highest F1 score on all the three subsets.

Imbalance ratio	100	50	10	5
BBN	58.0	56.3	52.5	49.3
LADE	59.5	57.6	53.3	51.5
PC	60.0	59.2	55.3	52.9
cRT	60.8	59.5	55.0	52.0
TDE	60.8	60.0	55.7	52.9
XE	**66.4**	63.9	56.3	51.9
xERM$_{PC}$	63.3	61.5	57.0	54.6
xERM$_{TDE}$	**66.4**	**64.8**	**59.3**	**55.5**

Table 4: Top-1 accuracy on CIFAR100-LT test set.

Imbalanced ratio	50	25	10	5
τ-Norm	59.6	58.2	56.2	54.6
LWS	60.6	59.2	57.0	55.0
PC	58.2	56.8	54.5	52.7
LADE	61.8	60.6	58.6	56.8
TDE	63.0	61.6	59.5	57.6
XE	**67.7**	65.2	61.4	58.0
xERM$_{PC}$	66.8	65.3	62.5	60.1
xERM$_{TDE}$	**67.7**	**66.0**	**63.5**	**60.9**

Table 5: Top-1 accuracy on ImageNet-LT test set.

Backbone	Acc	Recall			Precision			F1		
		Many	Med	Few	Many	Med	Few	Many	Med	Few
CIFAR100										
XE (PC)	52.6	60.3	51.9	44.4	59.6	51.1	44.4	60.0	51.5	44.4
TDE	52.6	60.4	51.7	44.4	59.5	51.0	44.5	60.0	51.4	44.5
LADE	53.9	58.7	53.8	47.8	60.2	54.5	47.1	59.4	54.1	47.4
xERM$_{TDE}$	55.1	62.8	54.5	46.7	61.7	53.9	48.1	62.3	54.2	47.4
xERM$_{PC}$	**55.3**	**60.9**	**56.0**	**48.0**	**63.7**	**54.3**	**48.3**	**62.3**	**55.1**	**48.1**
Places365										
XE (PC)	43.8	43.8	44.0	43.5	39.9	43.5	49.3	41.7	43.7	46.2
TDE	43.8	43.8	43.9	43.6	39.7	43.6	48.7	41.6	43.8	46.0
LADE	44.3	42.9	45.9	43.1	43.4	45.1	45.7	43.1	45.5	44.4
xERM$_{TDE}$	44.6	44.1	45.3	44.0	40.4	44.9	49.5	42.1	45.1	46.6
xERM$_{PC}$	**46.6**	**45.1**	**48.2**	**46.0**	**44.2**	**49.0**	**53.3**	**44.6**	**48.6**	**49.4**
ImageNet										
XE (PC)	56.5	64.5	53.8	43.2	59.8	55.1	50.6	62.1	54.4	46.6
TDE	56.5	64.4	53.8	43.7	60.2	55.2	49.8	62.2	54.5	46.6
LADE	57.9	62.6	55.7	52.2	62.4	56.5	52.9	62.5	56.1	52.5
xERM$_{TDE}$	58.9	66.5	56.4	46.2	62.1	57.8	**63.2**	64.2	57.1	49.4
xERM$_{PC}$	**60.2**	**64.8**	**58.2**	**53.8**	**64.9**	**58.3**	53.9	**64.8**	**58.2**	**53.8**

Table 6: Studies on the effectiveness of feature representation.

Evaluations on Imbalanced Test Set

Motivation. Most state-of-the-art methods assume the test sets are well-balanced (*i.e.*, IB=1) and leverage this prior to improve the performance, *i.e.*, playing a "prior flipping" game. As most long-tailed classification methods lift up the overall accuracy at the expense of head performance loss, evaluation on balanced test set alone cannot reveal their shortcoming. We considered a new setting where evaluations are conducted under long-tailed test set with different imbalance ratios. This setting is complementary to the balanced test for the unbiasedness evaluation, *i.e.*, performing well on both imbalanced and balanced test distributions.

In addition, in real-world applications, the test distributions are usually unknown to us, and the classes are often long-tailed distributed because of Zipf's law of nature, *e.g.*, cancer detection. Evaluation under long-tailed test set is closer to the wild. Note that this setting is different from Hong et al. (2021), where the test class prior is known during the training stage. In our setting, no side information about the test distribution prior is accessible to the models.

Settings. The model architectures are the same as the ones for balanced test and the test splits for imbalanced test are similarly long-tailed like the train set. We established the long-tailed test splits by downsampling the original well-balanced test set with various imbalanced ratios, which is the same as the training set construction.

Results. Table 4 and 5 show the top-1 accuracy on various long-tailed test splits. As expected, XE performs well when the test set is also imbalanced. Compared to XE, all the previous balanced methods enjoy performance gains on the well-balanced test set while achieves much lower accuracies on the imbalanced test splits. For example, in Table 4, when IB=100, same as the training set, the accuracies of all balanced models are at least 5.6% less than that of XE.

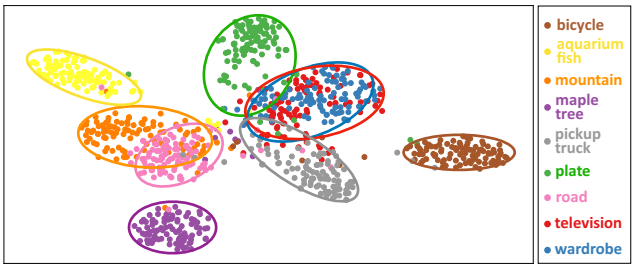
Note that our xERM outperforms all the other balanced methods in every imbalance setting, and achieves competitive results against TE method even when testing on extreme imbalanced ratio, *e.g.*, IB=100 on the CIFAR100 test split. When the imbalance ratio of the train set and test set are different, the performance of xERM is even higher than that of XE. These results demonstrate the advantage of xERM: *our proposed method is agnostic to both test distributions, i.e., it performs well on balanced and imbalanced test set.*

Evaluations on Feature Representation

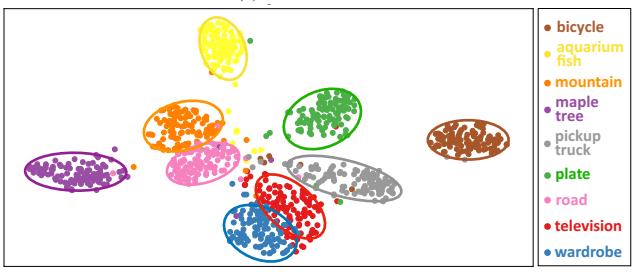
Settings. We conducted an empirical study on the quality of the feature representations. Specifically, we fixed the backbone for feature extraction, and re-trained the classifier head on a balanced dataset. Take CIFAR100-LT as an example. The models were first trained on the CIFAR100-LT dataset with IB=100. Then we fixed the backbone, and fine-tuned the classifiers on the balanced full CIFAR100 train set. Table 6 compares the performances on three benchmarks and we used ResNet-32, ResNet-50 and ResNeXt-50 as backbones for CIFAR100, Places365 and ImageNet, respectively.

Results on the empirical study. The classifiers built upon the xERM backbones consistently yield large improvement in terms of accuracy, recall and precision across the three datasets. xERM$_{PC}$ gains 2.7%, 2.8% and 3.7% on top-1 accuracy on CIFAR100, Places365 and ImageNet datasets, respectively. These results indicate that the xERM indeed produces better feature representation. Moreover, as expected, the performance gap between XE and TDE backbones is marginal. *This indicates that TDE does not improve the quality of feature representation, and it merely adjusts the biased decision boundary to achieve overall improvement.* Also, PC does not modify the training process, which leads to the same backbone as XE-training. As comparisons, our xERM achieves consistent improvements on top of both TDE and PC.

Results on t-SNE visualization. The t-SNE visualization (Van der Maaten and Hinton 2008) of embedding space



(a) t-SNE of TDE



(b) t-SNE of xERM_{TDE}

Figure 3: t-SNE visualizations on CIFAR100-LT. (a) visual features learned by TDE and (b) our xERM_{TDE}.

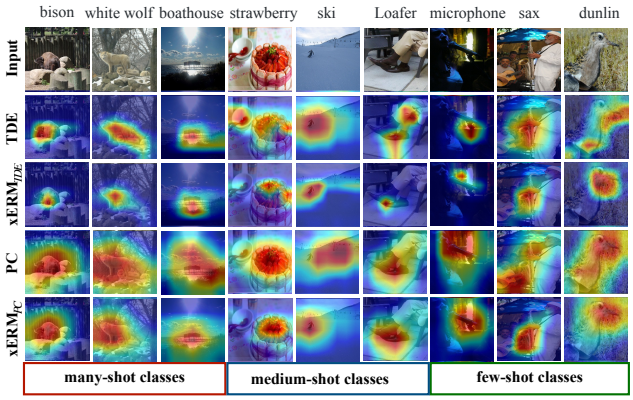


Figure 4: CAM visualizations on ImageNet-LT.

on CIFAR100-LT test set is plotted in Figure 3. We selected 9 classes and their class legends are listed from head (top) to tail (bottom). Compared with baseline TDE (Tang, Huang, and Zhang 2020), feature embedding of xERM_{TDE} is more compact and better separated, *e.g.*, "television" and "wardrobe" are overlapped in the feature embedding space of TDE (*cf.* Figure 3 (a)), but they are more separable in the feature embedding space of xERM_{TDE} (*cf.* Figure 3 (b)).

Results on CAM visualization. Figure 4 further visualized the activation maps using CAM (Zhou et al. 2016). Compared to TDE and PC, our xERM_{TDE} and xERM_{PC} yield more compact heatmap. For example, for "Loafer", our method pays more attention on shoes rather than the legs; in column "ski", our method focuses on the skier instead of the snow background. The visualization indicates that the predictions of xERM are less likely to be influenced by context region.

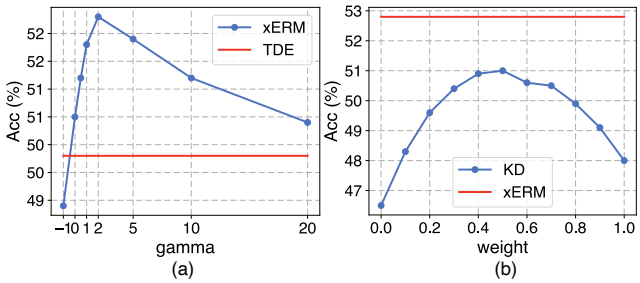


Figure 5: (a) Comparisons with a varying constant weight w. on CIFAR100-LT-IB-50 dataset. (b) Comparisons with a varying gamma γ in Eq. (9) on CIFAR100-LT-IB-50 dataset.

Ablations

Analysis of weights. Recall that w^{imba} and w^{ba} control the balance between the two losses. Readers may wonder whether the improvement comes from knowledge distillation, which shares similar formulation except for using constant weights $w \in [0, 1]$ and $1 - w$ for w^{ba} and w^{imba} to blend the KL divergence loss and cross-entropy loss. In particular, $w = 0$ denotes that model is only trained using the XE loss while $w = 1$ denotes self-distillation (Zhang et al. 2019). Figure 5 (a) illustrates the results of using constant weights ranging from 0 to 1. As shown in Figure 5 (a), knowledge distillation indeed can improve the performance compared to XE. However, it exhibits a large performance gap with xERM. This indicates that xERM is not a simple knowledge distillation approach.

Effect of the parameter γ. Recall that the hyper-parameter γ with a large absolute value would push w towards 0 or 1. In particular, $\gamma = 0$ is equivalent to a joint learning with $w^{\text{imba}} = w^{\text{ba}} = 0.5$, and $\gamma < 0$ denotes reversing the ratio in Eq. (9). Figure 5 (b) shows the influence of the value of gamma γ on the CIFAR100-LT-IB-50 dataset. Overall, xERM on top of TDE reaches its optimal accuracy with $\gamma = 2$, and consistently outperforms TDE with $\gamma > 0$. When γ is set as a negative value, the model underperforms TDE. This observation empirically verifies the correctness of Eq. (9).

Conclusion

Despite the overall accuracy improvement on the balanced test set, today's long-tailed models are still far from "improved" and exhibit the bias towards tail classes. In this work, we proposed xERM to train an unbiased test-agnostic model for long-tailed classification. Grounded by a causal theory, xERM removes the bias via minimizing the cross-domain empirical risk. Extensive experiments demonstrate that the strong improvement on both balanced and imbalanced test sets indeed comes from better feature representation rather than catering to the tail. The limitations of xERM are threefold: 1) it relies on the quality of the balanced model of the unseen balanced domain, 2) the counterfactual is currently only able to imagine one balanced distribution, and 3) our two-stage training pipeline involves two extra models, which increases the computation cost during training. We believe this work provides a new perspective to the unbiasedness of long-tailed classification and will inspire future works.

Acknowledgments

The authors would like to thank the reviewers for their comments that help improve the manuscript. This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2021-01-002).

References

Buda, M.; Maki, A.; and Mazurowski, M. A. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106: 249–259.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *NeurIPS*.

Chen, L.; Yan, X.; Xiao, J.; Zhang, H.; Pu, S.; and Zhuang, Y. 2020. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10800–10809.

Chen, L.; Zheng, Y.; Niu, Y.; Zhang, H.; and Xiao, J. 2021. Counterfactual samples synthesizing and training for robust visual question answering. *arXiv preprint arXiv:2110.01013*.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *CVPR*.

Drumnond, C. 2003. Class Imbalance and Cost Sensitivity: Why Undersampling beats Oversampling. In *ICML-KDD Workshop*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, Y.-Y.; Wu, J.; and Wei, X.-S. 2021. Distilling Virtual Examples for Long-tailed Recognition. In *CVPR*.

Heckman, J. J. 1979. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 153–161.

Hong, Y.; Han, S.; Choi, K.; Seo, S.; Kim, B.; and Chang, B. 2021. Disentangling Label Distribution for Long-tailed Visual Recognition. In *CVPR*.

Iscen, A.; Araujo, A.; Gong, B.; and Schmid, C. 2021. Class-Balanced Distillation for Long-Tailed Visual Recognition. arXiv:2104.05279.

Japkowicz, N.; and Stephen, S. 2002. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5): 429–449.

Jung, Y.; Tian, J.; and Bareinboim, E. 2020. Learning causal effects via weighted empirical risk minimization. *Advances in Neural Information Processing Systems*, 33: 12697–12709.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2020. Decoupling Representation and Classifier for Long-Tailed Recognition. In *ICLR*.

Keele, L. 2015. The statistics of causal inference: A view from political methodology. *Political Analysis*, 23(3): 313–335.

Khan, S. H.; Hayat, M.; Bennamoun, M.; Sohel, F. A.; and Togneri, R. 2017. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8): 3573–3587.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *ICCV*.

Liu, J.; Sun, Y.; Han, C.; Dou, Z.; and Li, W. 2020. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *CVPR*.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019. Large-Scale Long-Tailed Recognition in an Open World. In *CVPR*.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2021. Long-tail learning via logit adjustment. In *ICLR*.

Nam, J.; Cha, H.; Ahn, S.; Lee, J.; and Shin, J. 2020. Learning from Failure: Training Debiased Classifier from Biased Classifier. In *Advances in Neural Information Processing Systems*.

Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.-S.; and Wen, J.-R. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12700–12710.

Niu, Y.; and Zhang, H. 2021. Introspective Distillation for Robust Question Answering. *Advances in Neural Information Processing Systems*, 34.

Pearl, J. 2000. *Causality: models, reasoning and inference*, volume 29. Springer.

Pearl, J.; and Bareinboim, E. 2011. Transportability across studies: A formal approach. Technical report, CALIFORNIA UNIV LOS ANGELES DEPT OF COMPUTER SCIENCE.

Pearl, J.; Glymour, M.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Qi, J.; Niu, Y.; Huang, J.; and Zhang, H. 2020. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10860–10869.

Robins, J. M. 2001. Data, design, and background knowledge in etiologic inference. *Epidemiology*, 313–320.

Sarafianos, N.; Xu, X.; and Kakadiaris, I. A. 2018. Deep imbalanced attribute classification using visual attention aggregation. In *ECCV*.

Shen, L.; Lin, Z.; and Huang, Q. 2016. Relay backpropagation for effective learning of deep convolutional neural networks. In *ECCV*.

Tan, J.; Wang, C.; Li, B.; Li, Q.; Ouyang, W.; Yin, C.; and Yan, J. 2020. Equalization loss for long-tailed object recognition. In *CVPR*.

Tang, K.; Huang, J.; and Zhang, H. 2020. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *arXiv preprint arXiv:2009.12991*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The INaturalist Species Classification and Detection Dataset. In *CVPR*.

Wang, X.; Lian, L.; Miao, Z.; Liu, Z.; and Yu, S. X. 2020. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*.

Xiang, L.; Ding, G.; and Han, J. 2020. Learning From Multiple Experts: Self-paced Knowledge Distillation for Long-Tailed Classification. In *ECCV*.

Yang, X.; Zhang, H.; and Cai, J. 2021. Deconfounded image captioning: A causal retrospect. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yang, X.; Zhang, H.; Qi, G.; and Cai, J. 2021. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9847–9857.

Yue, Z.; Wang, T.; Sun, Q.; Hua, X.-S.; and Zhang, H. 2021. Counterfactual zero-shot and open-set visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15404–15414.

Zhang, L.; Song, J.; Gao, A.; Chen, J.; Bao, C.; and Ma, K. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *CVPR*.

Zhang, S.; Chen, C.; Hu, X.; and Peng, S. 2021a. Balanced Knowledge Distillation for Long-tailed Learning. *arXiv e-prints*.

Zhang, S.; Li, Z.; Yan, S.; He, X.; and Sun, J. 2021b. Distribution Alignment: A Unified Framework for Long-tail Visual Recognition. In *CVPR*.

Zhou, B.; Cui, Q.; Wei, X.-S.; and Chen, Z.-M. 2020. BBN: Bilateral-Branch Network with Cumulative Learning for Long-Tailed Visual Recognition. In *CVPR*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning Deep Features for Discriminative Localization. In *CVPR*.