Transfer Knowledge from Head to Tail: Uncertainty Calibration under Long-tailed Distribution (Appendix)

Jiahao Chen[1 2], Bing Su[1 2 †]

[1] Gaoling School of Artificial Intelligence, Renmin University of China

[2] Beijing Key Laboratory of Big Data Management and Analysis Methods

{nicelemon666, subingats}@gmail.com

1. The proof of Theorem 3.1

Following previous works [1, 3, 6, 8], we analyze the error bound of the importance weight strategy. We denote $p_k(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{p_k}, \boldsymbol{\Sigma}_{p_k})$ as the long-tailed distribution and $q_k(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{q_k}, \boldsymbol{\Sigma}_{q_k})$ as the ground truth balanced distribution, and $q_k^*(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{q_k^*}, \boldsymbol{\Sigma}_{q_k^*})$ as estimated distribution, where index k denotes the k^{th} class. For simplicity, we analyze the error bound of a given class k, and the error bound of other tail classes can also be analyzed following the same procedure. We denote importance weight $w_k(\boldsymbol{x}) = q_k(\boldsymbol{x})/p_k(\boldsymbol{x})$ and $w_k^*(\boldsymbol{x}) = q_k^*(\boldsymbol{x})/p_k(\boldsymbol{x})$. The unbiased error is:

$$\begin{aligned}
error &= |\mathbb{E}_{p_k}[w_k(\boldsymbol{x})\mathcal{L}(s(\boldsymbol{z}/T), y)] - \mathbb{E}_{p_k}[w_k^*(\boldsymbol{x})\mathcal{L}(s(\boldsymbol{z}/T), y)]| \\
&= |\mathbb{E}_{p_k}[w(\boldsymbol{x})\mathcal{L}(s(\boldsymbol{z}/T), y) - w_k^*(\boldsymbol{x})\mathcal{L}(s(\boldsymbol{z}/T), y)]| \\
&\leq \sqrt{\mathbb{E}_{p_k}[(w_k(\boldsymbol{x}) - w_k^*(\boldsymbol{x}))^2]\mathbb{E}_{p_k}[(\mathcal{L}(s(\boldsymbol{z}/T), y))^2]} \qquad \text{(Cachy-Schwarz Ineqaulity)} \\
&\leq \frac{1}{2}(\mathbb{E}_{p_k}[(w_k(\boldsymbol{x}) - w_k^*(\boldsymbol{x}))^2] + \mathbb{E}_{p_k}[(\mathcal{L}(s(\boldsymbol{z}/T), y))^2]) \quad \text{(AM/GM Inequality)}
\end{aligned} \tag{1}$$

As shown in Eq. (1), the unbiased error is sensitive to the term $\mathbb{E}_{p_k}[(w_k(\boldsymbol{x}) - w_k^*(\boldsymbol{x}))^2]$ since $\mathbb{E}_{p_k}[(\mathcal{L}(s(\boldsymbol{z}/T), y))^2])$ is determined. To better understand our method, we analyze the upper bound and lower bound for the first term and denote $\epsilon = \mathbb{E}_{p_k}[(w_k(\boldsymbol{x}) - w_k^*(\boldsymbol{x}))^2]$. The Eq. (2) shows that the upper bound of error, that is $d_2(q_k||p_k) + d_2(q_k^*||p_k)$. The formula $d_2(q||p)$ presents the exponential in base 2 of the Renyi-divergence [7] and is defined in Eq. (4) and Eq. (5).

$$\begin{aligned}
\epsilon &= \mathbb{E}_{p_k}[(w(\boldsymbol{x}_i) - w^*(\boldsymbol{x}_i))^2] \\
&= \mathbb{V}_{p_k}[w_k(\boldsymbol{x}) - w_k^*(\boldsymbol{x})] + (\mathbb{E}_{p_k}[w_k(\boldsymbol{x}) - w_k^*(\boldsymbol{x})])^2 \\
&= \mathbb{V}_{p_k}[w_k(\boldsymbol{x}) - w_k^*(\boldsymbol{x})] \qquad (\mathbb{E}_{p_k}[w_k(\boldsymbol{x})] = \mathbb{E}_{p_k}[w_k^*(\boldsymbol{x})] = 1) \\
&= \mathbb{V}_{p_k}[w_k(\boldsymbol{x})] + \mathbb{V}_{p_k}[w_k^*(\boldsymbol{x})] - 2Cov(w_k(\boldsymbol{x}), w_k^*(\boldsymbol{x})) \quad (Cov \text{ denotes covariance function.}) \\
&= d_2(q_k||p_k) + d_2(q_k^*||p_k) - 2Cov(w_k(\boldsymbol{x}), w_k^*(\boldsymbol{x})) - 2 \\
&= d_2(q_k||p_k) + d_2(q_k^*||p_k) - 2\mathbb{E}_{p_k}[w_k(\boldsymbol{x})w_k^*(\boldsymbol{x})] \\
&\leq d_2(q_k||p_k) + d_2(q_k^*||p_k)
\end{aligned} \tag{2}$$

We also analyze the lower bound of our strategy, which is shown in Eq. (3). This indicates that the error of our method

[†] Corresponding author.

will be larger than $(\sqrt{d_2(q_k||p_k)} - \sqrt{d_2(q_k^*||p_k)})^2$.

$$\begin{aligned}
\epsilon &= d_2(q_k||p_k) + d_2(q_k^*||p_k) - 2\mathbb{E}_{p_k}[w_k(\boldsymbol{x})w_k^*(\boldsymbol{x})] \\
&\geq d_2(q_k||p_k) + d_2(q_k^*||p_k) - 2\sqrt{\mathbb{E}_{p_k}[(w_k(\boldsymbol{x}))^2]\mathbb{E}_{p_k}[(w_k^*(\boldsymbol{x}))^2]} \\
&= d_2(q_k||p_k) + d_2(q_k^*||p_k) - 2\sqrt{(\mathbb{V}_{p_k}[w_k(\boldsymbol{x}))] + 1)(\mathbb{V}_{p_k}[w_k^*(\boldsymbol{x}))] + 1)} \\
&= d_2(q_k||p_k) + d_2(q_k^*||p_k) - 2\sqrt{d_2(q_k||p_k)(d_2(q_k^*||p_k)} \\
&= (\sqrt{d_2(q_k||p_k)} - \sqrt{d_2(q_k^*||p_k)})^2
\end{aligned} \tag{3}$$

$$D_\alpha(q||p) = \frac{\alpha}{2}(\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)^T[\alpha\boldsymbol{\Sigma}_p + (1-\alpha)\boldsymbol{\Sigma}_q]^{-1}(\boldsymbol{\mu}_q - \boldsymbol{\mu}_p) - \frac{1}{2(\alpha-1)}\ln\frac{|\alpha\boldsymbol{\Sigma}_p + (1-\alpha)\boldsymbol{\Sigma}_q|}{|\boldsymbol{\Sigma}_q|^{1-\alpha}|\boldsymbol{\Sigma}_p|^{\alpha}} \tag{4}$$

$$d_\alpha(q||p) = 2^{D_\alpha(q||p)} \tag{5}$$

Therefore, we draw the conclusion that $\epsilon \in [(\sqrt{d_2(q_k||p_k)} - \sqrt{d_2(q_k^*||p_k)})^2, d_2(q_k||p_k) + d_2(q_k^*||p_k)]$.

2. The calculation of τ_1 and τ_2

For simplicity, we constitute one-dimensional Gaussian distribution $p(x) = \mathcal{N}(x|\mu_a, \sigma_a^2)$ and $q(x) = \mathcal{N}(x|\mu_b, \sigma_b^2)$, where $\mu_a \neq \mu_b$ and $\sigma_a^2 < \sigma_b^2$. When $q(x) > p(x)$ we have $w(x) > 1$. As shown in Eq. (6), the formula equals to solve the inequality $h(x) = 2\sigma_b^2(x-\mu_a)^2 - 2\sigma_a^2(x-\mu_b)^2 + 4\sigma_a^2\sigma_b^2(\ln\sigma_a - \ln\sigma_b) > 0$. $h(x)$ is a quadratic function.

$$\begin{gathered}
q(x) > p(x) \\
\ln q(x) > \ln p(x) \\
\ln\sigma_a - \frac{(x-\mu_b)^2}{2\sigma_b^2} > \ln\sigma_b - \frac{(x-\mu_a)^2}{2\sigma_a^2} \\
4\sigma_a^2\sigma_b^2\ln\sigma_a - 2\sigma_a^2(x-\mu_b)^2 > 4\sigma_a^2\sigma_b^2\ln\sigma_b - 2\sigma_b^2(x-\mu_a)^2
\end{gathered} \tag{6}$$

The equation of $h(x) = 0$ have two solutions and denote as τ_1 and τ_2, $\tau_1 < \tau_2$, respectively. The condition for the inequality to hold is $x > \tau_2$ or $x < \tau_1$. Normally, the solution of quadratic function is $\frac{-B\pm\sqrt{B^2-4AC}}{2A}$. The coefficient of our function is shown in Eq. (7). Therefore, we get the value of τ_1 and τ_2 in Eq. (8).

$$\begin{aligned}
A &= 2\sigma_b^2 - 2\sigma_a^2 \\
B &= 4\mu_b\sigma_a^2 - 4\mu_a\sigma_b^2 \\
C &= 2\mu_a^2\sigma_b^2 - 2\mu_b^2\sigma_a^2 - 4\sigma_a^2\sigma_b^2(\ln\sigma_b - \ln\sigma_a)
\end{aligned} \tag{7}$$

$$\begin{aligned}
\tau_1 &= \frac{\mu_a\sigma_b^2 - \mu_b\sigma_a^2 - \sigma_a\sigma_b\sqrt{(\mu_a-\mu_b)^2 + (\sigma_b^2-\sigma_a^2)(\ln\sigma_b^2 - \ln\sigma_a^2)}}{\sigma_b^2 - \sigma_a^2} \\
\tau_2 &= \frac{\mu_a\sigma_b^2 - \mu_b\sigma_a^2 + \sigma_a\sigma_b\sqrt{(\mu_a-\mu_b)^2 + (\sigma_b^2-\sigma_a^2)(\ln\sigma_b^2 - \ln\sigma_a^2)}}{\sigma_b^2 - \sigma_a^2}
\end{aligned} \tag{8}$$

3. Training procedures

Our calibration method consists of four procedures. 1. Estimate the feature distribution of each class on the validation set: the effect is to obtain the statistics for both head and tail classes for similarity calculation and knowledge transfer. 2. Calculate attentions between head and tail classes based on the Wasserstein distance between their distributions: the effect is to determine how much knowledge is transferred from each head class to a tail class in a principled manner. 3. Estimate importance weights with the calibrated distributions: the effect is to compensate for tail classes by reweighting their samples. 4. Learn the temperature T with the importance weights: the effect is to scale prediction confidence scores for calibration under long-tailed distribution.

CIFAR-100	ResNet-32	DenseNet-40	VGG-19
ours(W-dist)	**1.50**	**2.37**	**1.99**
Re-weighting	1.90	2.45	2.18
Re-sampling	1.83	2.55	2.00

Table 1. The ECE(%) on CIFAR-100-LT with IF=10.

4. Dataset and training strategy

Dataset. The distribution of CIFAR-10-LT, MNIST-LT, CIFAR-100-LT, and ImageNet-LT, are shown in Fig. 1, Fig. 2, Fig. 3, and Fig. 4, respectively. The training set and validation set follows the long-tailed distribution while the test set is not. For CIFAR-100-LT and ImageNet-LT, the local distribution of the validation set and the training set exists a little difference. Since we split data randomly and such a phenomenon is rational.

Training strategy. For CIFAR-10-LT and CIFAR-100-LT, we use ResNet-32 as our backbone following [2]. We use the SGD optimizer and set the initial learning rate to 0.1. The model has trained a total of 200 epochs. The first five epochs are trained with the linear warm-up [4] learning rate schedule. The learning rate drops by 0.1 at epoch 160 and epoch 180, respectively. We follow the most popular setting to set the batch size, the momentum, and the weight decay to 128, 0.9, and 5×10^{-4}, respectively. For MNIST-LT, we use LeNet-5 as the backbone. We use an SGD optimizer and set the initial learning rate to 0.1. The model has trained a total of 100 epochs. The learning rate drops by 0.1 at epoch 60. We follow the most popular setting to set the batch size, the momentum, and the weight decay to 256, 0.9, and 5×10^{-4}, respectively. For ImageNet-LT, we use ResNet-50 as backbones and adopt the cosine learning rate schedule [5] that gradually decays from 0.1 to 0 in the first stage. The model has trained a total of 180 epochs. We follow the most popular setting to set the batch size, the momentum, and the weight decay to 256, 0.9, and 5×10^{-4}, respectively.

5. More ablation studies

We compare our method with the inverse frequency sampling (resampling) and loss weighting (reweighting) in Tab.1. Since there exists some randomness of instance selection in the re-sampling method, the results of re-weighting and re-sampling are slightly different. Our method outperforms these model-agnostic baselines. It relies on the features extracted by the model, so the estimated importance weights can better reflect the bias of the model.

We employ the Wasserstein distance (W-dist) because it incorporates second-order statistics that can be important because calibration aims to make the prediction uncertainty more precise. W-dist between two Gaussian distributions has a closed form. As shown in Tab.2, W-dist outperforms L_2 distance and cosine similarity that does not consider second-order statistics.

We split the class id of 0-2 as head classes on CIFAR-10-LT, which follows MiSLAS [9]. Tab.2 shows the results of different split strategies. We also evaluate the split by utilizing the similarity of all classes, whose results are worse than ours. This is because the distributions of tail classes are not reliable due to the few samples, thus transferring such knowledge is harmful.

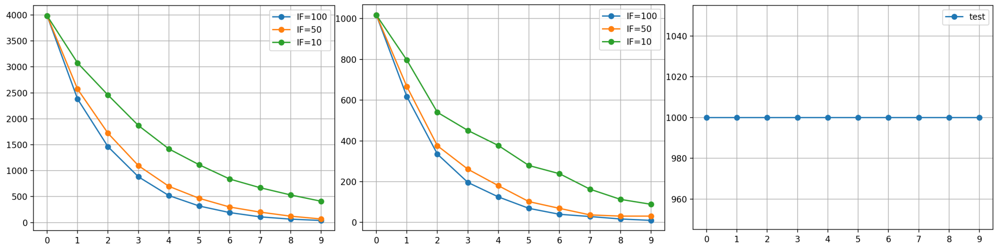
We also demonstrate the effectiveness of our attention mechanism. As shown in Tab.2, our attention mechanism outperforms the simple normalization strategy, where the inverse W-dist values are directly normalized. We use this attention mechanism because its effectiveness has been demonstrated in many areas and the softmax function enhances the knowledge transfer from closer header classes.

6. More evaluation metrics

We evaluate our method with different evaluation metrics. The results are shown from Tab. 3 to Tab. 14. Ours1 and ours2 denote our method with $\alpha = 0.998$ and $\alpha = 0.995$, respectively. For the SCE metric and ACE metric, our method achieves competitive results. The accuracy table shows that our method will preserve the model's accuracy.

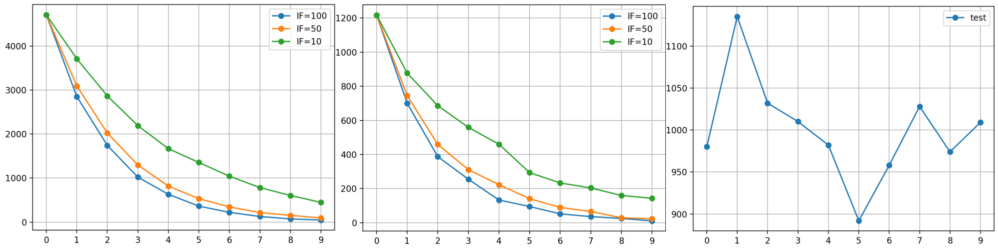
CIFAR-10	IF=100	IF=50	IF=10
ours(W-dist)	9.84	**3.99**	**1.00**
class id 0-1	**9.14**	4.29	1.07
class id 0-5	10.22	4.91	1.36
class id 0-8	12.21	6.48	2.13
similarity of all classes	9.91	4.05	1.17
L_2 distance	10.15	4.18	1.22
cosine similarity	10.41	4.53	1.46
normalization	10.39	4.51	1.48

Table 2. The ECE(%) on CIFAR-10-LT.



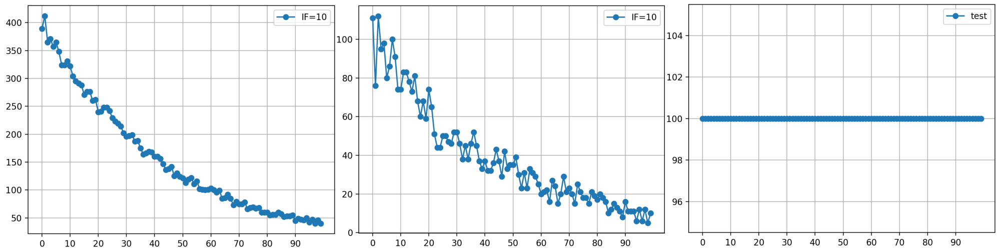
(a) Training set (b) Validation set (c) Test set

Figure 1. The distribution of long-tailed CIFAR-10-LT. The horizontal axis represents the class index and the vertical axis represents the number of instances.



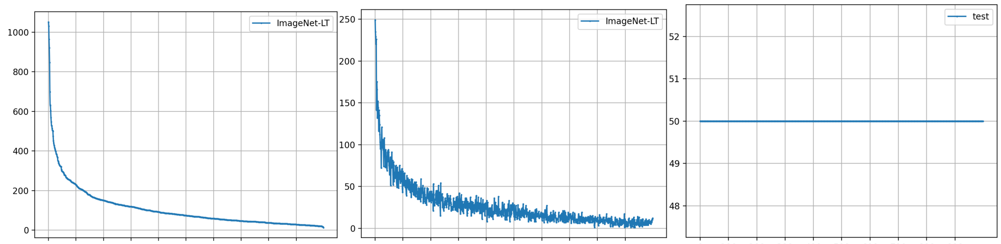
(a) Training set (b) Validation set (c) Test set

Figure 2. The distribution of long-tailed MNIST-LT. The horizontal axis represents the class index and the vertical axis represents the number of instances.



(a) Training set (b) Validation set (c) Test set

Figure 3. The distribution of long-tailed CIFAR-100-LT. The horizontal axis represents the class index and the vertical axis represents the number of instances.



(a) Training set (b) Validation set (c) Test set

Figure 4. The distribution of long-tailed ImageNet-LT. The indices are sorted by the number of instances per class. The horizontal axis represents the class index and the vertical axis represents the number of instances.

IF	**Dataset**	**Method**									
		Base	TS	ETS	TS-IR	IR	IROvA	SBC	GPC	Ours1	Ours2
IF=100	CIFAR-10	5.11	4.21	4.20	4.03	4.09	4.22	4.09	4.33	4.08	**3.97**
	CIFAR-10.1	6.49	4.97	4.96	4.82	5.01	5.02	4.93	5.07	4.73	**4.49**
	CIFAR-10.1-C	12.63	10.21	10.2	10.34	10.31	10.21	10.07	10.39	9.65	**9.00**
	CIFAR-F	6.61	4.95	4.94	5.01	4.99	4.96	4.96	5.05	4.70	**4.47**
IF=50	CIFAR-10	4.08	3.36	3.36	**3.19**	3.26	3.35	3.20	3.48	3.27	3.29
	CIFAR-10.1	5.29	4.03	4.06	3.81	4.02	4.04	3.88	4.15	3.82	**3.76**
	CIFAR-10.1-C	12.18	9.56	9.69	9.64	9.74	9.50	9.40	9.81	8.67	**8.14**
	CIFAR-F	5.87	4.32	4.37	4.38	4.51	4.32	4.30	4.49	4.07	**4.04**
IF=10	CIFAR-10	1.93	1.31	1.29	1.32	1.32	1.30	**1.26**	1.47	1.27	1.34
	CIFAR-10.1	3.17	1.97	1.88	2.12	2.17	1.91	2.04	2.25	**1.84**	1.83
	CIFAR-10.1-C	10.91	8.51	8.25	8.76	8.58	8.35	8.48	8.91	8.09	**7.60**
	CIFAR-F	4.56	3.22	3.14	3.32	3.35	3.23	3.21	3.37	3.11	**3.09**

Table 3. The SCE (%) on CIFAR-10-LT.

IF	**Dataset**	**Method**									
		Base	TS	ETS	TS-IR	IR	IROvA	SBC	GPC	Ours1	Ours2
IF=100	CIFAR-10	4.99	4.16	4.15	4.02	4.07	4.15	4.05	4.30	4.04	**3.99**
	CIFAR-10.1	6.22	4.89	4.89	4.8	4.85	4.96	4.78	5.00	4.65	**4.45**
	CIFAR-10.1-C	12.44	10.16	10.15	10.27	10.24	10.14	10.03	10.33	9.62	**9.01**
	CIFAR-F	6.37	4.87	4.87	4.87	4.84	4.86	4.83	4.93	4.66	**4.51**
IF=50	CIFAR-10	3.96	3.33	3.35	3.2	3.24	3.33	**3.17**	3.47	3.31	3.37
	CIFAR-10.1	4.99	4.02	4.06	3.84	3.93	3.96	**3.83**	4.14	**3.83**	**3.83**
	CIFAR-10.1-C	11.97	9.5	9.63	9.59	9.66	9.44	9.36	9.76	8.67	**8.17**
	CIFAR-F	5.62	4.28	4.33	4.31	4.43	4.28	4.26	4.38	4.14	**4.13**
IF=10	CIFAR-10	1.76	1.31	1.28	1.29	1.28	1.30	**1.23**	1.42	1.27	1.35
	CIFAR-10.1	2.82	1.81	1.76	1.91	1.89	1.78	1.87	1.94	**1.73**	1.78
	CIFAR-10.1-C	10.71	8.47	8.23	8.68	8.52	8.35	8.44	8.83	8.08	**7.61**
	CIFAR-F	4.29	3.12	3.07	3.14	3.12	3.08	3.06	3.18	3.04	**3.02**

Table 4. The ACE (%) on CIFAR-10-LT.

IF	**Dataset**	**Method**									
		Base	TS	ETS	TS-IR	IR	IROvA	SBC	GPC	Ours1	Ours2
IF=100	CIFAR-10	69.38	69.38	69.38	70.55	70.27	68.86	69.89	69.28	69.38	69.38
	CIFAR-10.1	59.80	59.80	59.80	60.90	60.50	59.30	60.35	59.75	59.80	59.80
	CIFAR-10.1-C	27.55	27.55	27.55	27.74	27.70	27.40	27.86	27.54	27.55	27.55
	CIFAR-F	57.79	57.79	57.79	58.45	58.56	57.29	58.27	57.76	57.79	57.79
IF=50	CIFAR-10	74.68	74.68	74.68	74.92	74.65	74.22	74.92	74.7	74.68	74.68
	CIFAR-10.1	66.10	66.10	66.10	65.80	66.00	65.40	66.45	66.10	66.10	66.10
	CIFAR-10.1-C	29.06	29.06	29.06	28.75	28.71	28.90	29.08	29.05	29.06	29.06
	CIFAR-F	61.51	61.51	61.51	61.3	61.21	61.06	61.49	61.49	61.51	61.51
IF=10	CIFAR-10	86.10	86.10	86.10	86.36	86.09	85.86	86.14	86.13	86.10	86.10
	CIFAR-10.1	77.75	77.75	77.75	77.9	77.75	77.45	77.9	77.75	77.75	77.75
	CIFAR-10.1-C	33.45	33.45	33.45	33.79	33.66	33.34	33.60	33.45	33.45	33.45
	CIFAR-F	67.84	67.84	67.84	68.24	68.06	67.65	68.12	67.83	67.84	67.84

Table 5. The Accuracy (%) on CIFAR-10-LT.

IF	**Dataset**	**Method**									
		Base	TS	ETS	TS-IR	IR	IROvA	SBC	GPC	Ours1	Ours2
IF=100	MNIST	**0.82**	0.85	**0.82**	0.97	0.97	0.85	0.92	0.83	0.87	0.90
	SVHN	4.77	3.68	4.06	7.25	7.35	4.59	4.30	4.39	3.61	**3.50**
	USPS	3.94	3.34	3.66	4.23	4.24	3.32	3.83	3.54	3.25	**3.20**
	Digital-S	7.99	6.24	7.16	8.07	8.11	6.91	6.97	7.14	6.00	**5.62**
IF=50	MNIST	**0.43**	0.44	**0.43**	0.49	0.49	**0.43**	0.50	**0.43**	0.45	0.47
	SVHN	3.08	3.22	3.07	6.33	6.71	**2.87**	5.65	3.08	3.28	3.43
	USPS	3.39	3.24	3.55	4.23	4.27	3.31	4.00	3.39	**3.20**	3.29
	Digital-S	5.11	4.59	4.48	8.05	8.20	4.55	7.24	5.11	4.45	**4.18**
IF=10	MNIST	**0.20**	0.22	0.22	0.21	0.22	0.21	0.24	0.22	0.23	0.23
	SVHN	**3.52**	3.73	3.76	6.21	5.51	3.56	3.77	3.58	3.80	3.85
	USPS	**2.91**	3.07	3.09	3.27	3.29	3.15	3.30	3.11	3.18	3.31
	Digital-S	4.86	4.27	4.25	6.02	5.69	4.35	4.81	4.43	**4.23**	**4.23**

Table 6. The SCE (%) on MNIST-LT.

IF	**Dataset**	**Method**									
		Base	TS	ETS	TS-IR	IR	IROvA	SBC	GPC	Ours1	Ours2
IF=100	MNIST	**0.78**	0.85	0.82	0.92	0.93	0.83	0.92	0.82	0.86	0.89
	SVHN	4.84	3.81	4.24	7.31	7.37	4.54	4.55	4.48	3.73	**3.65**
	USPS	3.82	3.28	3.56	4.11	4.02	3.38	3.69	3.43	3.21	**3.10**
	Digital-S	7.84	6.29	7.09	7.99	8.00	6.85	7.07	7.08	6.06	**5.79**
IF=50	MNIST	**0.40**	0.43	0.41	0.45	0.45	0.41	0.48	**0.40**	0.44	0.46
	SVHN	3.21	3.51	3.22	6.19	6.60	**3.12**	5.65	3.21	3.63	3.90
	USPS	3.26	3.15	3.37	4.01	4.01	3.27	3.89	3.26	**3.12**	3.18
	Digital-S	5.05	4.58	5.38	7.69	7.83	4.66	7.04	5.05	4.43	**4.31**
IF=10	MNIST	**0.17**	0.19	0.19	0.19	0.19	0.18	0.20	0.18	0.19	0.20
	SVHN	**3.79**	4.02	4.03	6.26	5.58	3.88	4.03	3.83	4.08	4.16
	USPS	**2.82**	3.03	3.05	3.05	3.10	3.06	3.09	2.89	3.10	3.22
	Digital-S	4.84	4.43	4.42	5.86	5.55	4.40	4.76	4.56	**4.37**	**4.37**

Table 7. The ACE (%) on MNIST-LT.

IF	**Dataset**	**Method**									
		Base	TS	ETS	TS-IR	IR	IROvA	SBC	GPC	Ours1	Ours2
IF=100	MNIST	95.12	95.12	95.12	94.69	94.57	94.92	94.84	95.17	95.12	95.12
	SVHN	26.23	26.23	26.23	22.85	23.39	25.74	26.74	26.27	26.23	26.23
	USPS	72.99	72.99	72.99	71.79	71.79	72.24	72.24	72.79	72.99	72.99
	Digital-S	30.88	30.88	30.88	28.84	28.84	30.13	30.75	30.81	30.88	30.88
IF=50	MNIST	97.44	97.44	97.44	97.49	97.5	97.34	97.27	97.44	97.44	97.44
	SVHN	36.26	36.26	36.26	32.16	31.79	35.8	30.13	36.27	36.26	36.26
	USPS	76.08	76.08	76.08	75.73	75.63	75.83	75.98	76.08	76.08	76.08
	Digital-S	41.58	41.58	41.58	39.26	38.8	41.13	38.36	41.62	41.58	41.58
IF=10	MNIST	98.57	98.57	98.57	98.38	98.35	98.56	98.50	98.56	98.57	98.57
	SVHN	35.71	35.71	35.71	29.69	30.6	33.49	35.57	35.76	35.71	35.71
	USPS	79.67	79.67	79.67	78.62	78.52	79.47	79.67	79.72	79.67	79.67
	Digital-S	43.38	43.38	43.38	40.26	41.03	41.81	42.49	43.5	43.38	43.38

Table 8. The Accuracy (%) on MNIST-LT.

Model	**Dataset**	**Method**									
		Base	TS	ETS	TS-IR	IR	IROvA	SBC	GPC	Ours1	Ours2
ResNet-32	CIFAR-100	0.52	0.33	**0.32**	0.39	0.40	0.34	0.38	0.33	0.33	**0.32**
DenseNet-40	CIFAR-100	0.44	**0.33**	**0.33**	0.37	0.37	**0.33**	0.35	0.34	**0.33**	**0.33**
VGG-19	CIFAR-100	0.64	0.28	0.28	0.28	0.29	0.29	0.36	0.29	**0.27**	0.28

Table 9. The SCE (%) on CIFAR-100-LT.

Model	**Dataset**	**Method**									
		Base	TS	ETS	TS-IR	IR	IROvA	SBC	GPC	Ours1	Ours2
ResNet-32	CIFAR-100	0.37	**0.28**	0.29	0.35	0.35	0.29	0.35	0.29	**0.28**	**0.28**
DenseNet-40	CIFAR-100	0.33	**0.28**	**0.28**	0.31	0.31	**0.28**	0.30	0.29	**0.28**	**0.28**
VGG-19	CIFAR-100	0.39	0.26	0.26	0.35	0.34	0.27	0.33	0.27	**0.26**	0.27

Table 10. The ACE (%) of CIFAR-100-LT.

Model	**Dataset**	**Method**									
		Base	TS	ETS	TS-IR	IR	IROvA	SBC	GPC	Ours1	Ours2
ResNet-32	CIFAR-100	56.13	56.13	56.13	54.67	54.89	55.98	55.52	55.85	56.13	56.13
DenseNet-40	CIFAR-100	60.39	60.39	60.39	59.74	59.5	60.25	60.40	60.41	60.39	60.39
VGG-19	CIFAR-100	56.06	56.06	56.06	54.4	54.8	56.01	55.63	55.97	56.06	56.06

Table 11. The Accuracy (%) on CIFAR-100-LT.

Model	**Dataset**	**Method**									
		Base	TS	ETS	TS-IR	IR	IROvA	SBC	GPC	Ours1	Ours2
ResNet-50	ImageNet	0.051	**0.049**	**0.049**	0.053	0.053	0.05	0.052	**0.049**	**0.049**	**0.049**

Table 12. The SCE (%) of ImageNet-LT.

Model	**Dataset**	**Method**									
		Base	TS	ETS	TS-IR	IR	IROvA	SBC	GPC	Ours1	Ours2
ResNet-50	ImageNet	0.042	**0.041**	0.042	0.043	0.043	**0.041**	0.043	**0.041**	**0.041**	**0.041**

Table 13. The ACE (%) on ImageNet-LT.

Model	**Dataset**	**Method**									
		Base	TS	ETS	TS-IR	IR	IROvA	SBC	GPC	Ours1	Ours2
ResNet-50	ImageNet	48.68	48.68	48.68	48.11	48.11	48.61	48.68	48.68	48.68	48.68

Table 14. The Accuracy (%) on ImageNet-LT.

References

[1] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. *Advances in neural information processing systems*, 23, 2010. 1

[2] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 3

[3] Yunye Gong, Xiao Lin, Yi Yao, Thomas G Dietterich, Ajay Divakaran, and Melinda Gervasio. Confidence calibration for domain generalization under covariate shift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8958–8967, 2021. 1

[4] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 3

[5] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 3

[6] Anusri Pampari and Stefano Ermon. Unsupervised calibration under covariate shift. *arXiv preprint arXiv:2006.16405*, 2020. 1

[7] Alfréd Rényi et al. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1. Berkeley, California, USA, 1961. 1

[8] Ximei Wang, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable calibration with lower bias and variance in domain adaptation. *Advances in Neural Information Processing Systems*, 33:19212–19223, 2020. 1

[9] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16489–16498, 2021. 3