Supplementary Material for: Balanced Contrastive Learning for Long-Tailed Visual Recognition

Jianggang Zhu[1,2]*, Zheng Wang[1,2]*, Jingjing Chen[1,2]†, Yi-Ping Phoebe Chen[3] and Yu-Gang Jiang[1,2]
[1]Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University
[2]Shanghai Collaborative Innovation Center on Intelligent Visual Computing
[3]Department of Computer Science and Information Technology, La Trobe University
{jgzhu20,zhengwang17,chengjingjing,ygj}@fudan.edu.cn, phoebe.chen@latrobe.edu.au

A. Proof of Theorem 2 and Theorem 3

In this section, we will proof Theorem 2 and Theorem 3 proposed in section 3.2. The main idea of the proof is to decouple the class-specific batch-wise loss as attraction term and repulsion term as in [2]. First, we will show the spontaneous appearance of variability collapse as the training process in attraction term. When this condition holds, we find that to minimize the loss, the solution of the model spontaneously satisfies the simplex configuration.

Before the detailed derivation, recall the main notions and definitions of this paper:

- $h, K, N \in \mathbb{N}$
- $\mathcal{Z} = \mathbb{R}^h$
- $\mathcal{Y} = [K] = \{1, 2, \ldots, K\}$

Definition 1 (Supervise contrastive loss) For an instance x_i of representation z_i in a batch B, supervised contrastive loss has the following expression:

$$\mathcal{L}_i = -\frac{1}{|B_y| - 1} \sum_{p \in B_y \setminus \{i\}} \log \frac{\exp(z_i \cdot z_p)}{\sum\limits_{k \in B \setminus \{i\}} \exp(z_i \cdot z_k)} \tag{1}$$

Definition 2 (Balanced contrastive loss) For an instance x_i of representation z_i in a batch B, balanced contrastive loss has the following expression:

$$\mathcal{L}_i = -\frac{1}{|B_y| - 1} \sum_{p \in B_y \setminus \{i\}} \log \frac{\exp(z_i \cdot z_p)}{\sum\limits_{j \in \mathcal{Y}_B} \frac{1}{|B_j|} \sum\limits_{k \in B_j} \exp(z_i \cdot z_k)} \tag{2}$$

Definition 3 (Class-specific batch-wise loss)

$$\mathcal{L}(Z; Y, B, y) = \begin{cases} \sum_{i \in B_y} \mathcal{L}_i & if\ |B_y| > 1 \\ 0 & else \end{cases} \tag{3}$$

Definition 4 (Regular simplex) A set of points $\zeta_1, \ldots, \zeta_K \in \mathbb{R}^h$ form the vertices of a regular simplex inscribed in the hypersphere of radius $\rho > 0$, if and only if the following conditions hold:

(1) $\sum_{i \in [K]} \zeta_i = 0$

(2) $\|\zeta_i\| = \rho$, for $i \in [K]$

*Indicates equal contribution.
†Jingjing-Chen is the corresponding author.

(3) $\exists d \in \mathbb{R} : d = \langle \zeta_i, \zeta_j \rangle$ for $1 \leq i < j \leq K$

where B_y and $\mathcal{Y}_B$ are subsets of B and $\mathcal{Y}$, respectively. Note that the $|B_j|$ term in above equations will minus one when the positive class is averaged. Here, we omit hyper parameter temperature τ and $\langle \cdot \rangle$ for the inner product operation. Additionally, we default to $K \geq h + 1$ and assume $\|z_i\|_2 = 1$.

Proof of Theorem 2 First we rewrite class-specific batch-wise loss as following form:

$$\begin{aligned}
\mathcal{L}_{BCL}(Z; Y, B, y) &= \sum_{i \in B_y} -\frac{1}{|B_y| - 1} \sum_{p \in B_y \setminus \{i\}} \log \frac{\exp(z_i \cdot z_p)}{\sum\limits_{j \in \mathcal{Y}_B} \frac{1}{|B_j|} \sum\limits_{k \in B_j} \exp(z_i \cdot z_k)} \\
&= \sum_{i \in B_y} \log \left(\frac{\sum\limits_{j \in \mathcal{Y}_B} \frac{1}{|B_j|} \sum\limits_{k \in B_j} \exp(z_i \cdot z_k)}{\prod\limits_{p \in B_y \setminus \{i\}} \exp(z_i, z_p)^{1/|B_y|-1}} \right) \\
&= \sum_{i \in B_y} \log \left(\frac{\sum\limits_{j \in \mathcal{Y}_B} \frac{1}{|B_j|} \sum\limits_{k \in B_j} \exp(z_i \cdot z_k)}{\exp(\frac{1}{|B_y|-1} \sum\limits_{p \in B_y \setminus \{i\}} z_i \cdot z_p)} \right)
\end{aligned} \tag{4}$$

The key idea is to divide the sum in the numerator into positives and negatives. Since the exponential function is convex, by applying Jensen's inequality, we have

$$\begin{aligned}
\frac{1}{|B_y| - 1} \sum_{k \in B_y \setminus \{i\}} \exp(z_i \cdot z_k) &\overset{(Q1)}{\geq} \exp \left(\frac{1}{|B_y| - 1} \sum_{k \in B_y \setminus \{i\}} z_i \cdot z_k \right) \\
\frac{1}{|B_j|} \sum_{\substack{k \in B_j \\ j \neq y}} \exp(z_i \cdot z_k) &\overset{(Q2)}{\geq} \exp \left(\frac{1}{|B_j|} \sum_{k \in B_j} z_i \cdot z_k \right)
\end{aligned} \tag{5}$$

The equality is attained if and only if:

(Q1) There is $C_i(B, y)$ such that $\forall k \in B_y \setminus \{i\}$ all inner products $z_i \cdot z_k = C_i(B, y)$ are equal.

(Q2) There is $D_i(B, y, j)$ such that $\forall k \in B_{j, j \neq y}$ all inner products $z_i \cdot z_k = D_i(B, y, j)$ are equal.

Thus, the sum in the numerator can be written as follows

$$\sum_{j \in \mathcal{Y}_B} \frac{1}{|B_j|} \sum_{j \in B_j} \exp(z_i \cdot z_k) \geq \exp \left(\frac{1}{|B_y| - 1} \sum_{p \in B_y \setminus \{i\}} z_i \cdot z_p \right) + \sum_{\substack{j \in \mathcal{Y}_B \\ j \neq y}} \exp \left(\frac{1}{|B_j|} \sum_{k \in B_j} z_i \cdot z_k \right) \tag{6}$$

By leverage Jensen's inequality again on the latter term, resulting in

$$\sum_{\substack{j \in \mathcal{Y}_B \\ j \neq y}} \exp \left(\frac{1}{|B_j|} \sum_{k \in B_j} z_i \cdot z_k \right) \overset{(Q3)}{\geq} (|\mathcal{Y}_B| - 1) \exp \left(\frac{1}{|\mathcal{Y}_B| - 1} \sum_{\substack{j \in \mathcal{Y}_B \\ j \neq y}} \frac{1}{|B_j|} \sum_{k \in B_j} z_i \cdot z_k \right) \tag{7}$$

Here, the equality is attained if and only if

(Q3) There is $E_i(B, y)$ such that $\forall j \in \mathcal{Y}_{B, j \neq y}, \forall k \in B_j$ all inner products $z_i \cdot z_k = E_i(B, y)$ are equal.

Thus, for a specific mini-batch, Eq. 4 can be written as

$$\mathcal{L}_{BCL}(Z; Y, B, y) \geq \sum_{i \in B_y} \log \left(1 + (|\mathcal{Y}_B| - 1) \exp \left(\underbrace{\frac{1}{|\mathcal{Y}_B| - 1} \sum_{q \in \mathcal{Y}_B \setminus \{y\}} \frac{1}{|B_q|} \sum_{k \in B_q} z_i \cdot z_k}_{repulsion\ term} \underbrace{- \frac{1}{|B_y| - 1} \sum_{j \in B_y \setminus \{i\}} z_i \cdot z_j}_{attraction\ term} \right) \right) \tag{8}$$

which ends the proof of Theorem 2. Here, the equality is attained if and only if conditions (Q1) and (Q3) hold for every $i \in B_y$. Additionally, constants $C_i(B, y)$ and $E_i(B, y)$ only depend on the batch B and the label y.

Proof of Theorem 3 On the basis of theorem 2, we assume $\mathcal{Y}_B = \mathcal{Y}$ for every batch B. For simplicity, we rewrite the two terms of the exponential function in Theorem 2 as the following form

$$\begin{aligned} S(Z;Y,B,y) &= S_{att}(Z;Y,B,y) + S_{rep}(Z;Y,B,y) \\ S_{att}(Z;Y,B,y) &= -\frac{1}{|B_y|-1} \sum_{j \in B_y \setminus \{i\}} z_i \cdot z_j \\ S_{rep}(Z;Y,B,y) &= \frac{1}{|\mathcal{Y}|-1} \sum_{q \in \mathcal{Y} \setminus \{y\}} \frac{1}{|B_q|} \sum_{k \in B_q} z_i \cdot z_k \end{aligned} \tag{9}$$

Regroup the addends, we can obtain the following formulation

$$\begin{aligned} \mathcal{L}_{BCL}(Z;Y) &= \sum_{B \in \mathcal{B}} \sum_{y \in \mathcal{Y}} \mathcal{L}_{BCL}(Z;Y,B,y) \\ &\geq \sum_{B \in \mathcal{B}} \sum_{y \in \mathcal{Y}} \sum_{i \in B_y} \log\left(1 + (|\mathcal{Y}|-1)\exp(S(Z;Y,B,y))\right) \end{aligned} \tag{10}$$

Let $\alpha > 0$, and $f : \mathbb{R} \to \mathbb{R}, x \to \log(1 + \alpha \exp(x))$. It is easy to verify that the function f is smooth with second derivative and convex. According to Jensen's inequality, we obtain the lower bound as follows

$$\mathcal{L}_{BCL}(Z;Y) \overset{(Q4)}{\geq} |\mathcal{D}| \log\left(1 + (|\mathcal{Y}|-1)\exp\left(\sum_{B \in \mathcal{B}} \sum_{y \in \mathcal{Y}} \sum_{i \in B_y} S(Z;Y,B,y)\right)\right) \tag{11}$$

where $\mathcal{D}$ denotes the dataset, the equality is attained if and only if:

(Q4) There is constant θ such that $\forall B \in \mathcal{B}, \forall y \in \mathcal{Y}$ and $\forall i \in B_y$, the values of $S(Z;Y,B,y) = \theta$ agree.

Next we derive the sum of attraction terms. For every $Y \in \mathcal{Y}^N$ and every $Z \in \mathcal{Z}^N$, using the Cauchy-Schwarz inequality and the assumption that $\mathcal{Z}$ is a unit hypersphere, we have

$$\begin{aligned} \sum_{i \in B_y} S_{att}(Z;Y,B,y) &= -\frac{1}{|B_y|-1} \sum_{i \in B_y} \sum_{j \in B_y \setminus \{i\}} z_i \cdot z_j \\ &\overset{(Q5)}{\geq} -|B_y| \times \frac{1}{|B_y|(|B_y|-1)} \sum_{i \in B_y} \sum_{j \in B_y \setminus \{i\}} \|z_i\|\|z_j\| = -|B_y| \end{aligned} \tag{12}$$

Since the z_i and z_j are on a hypersphere, this implies the condition of equality is equivalent to $z_i = z_j$.

(Q5) For every $n, m \in [N]$, $y_n = y_m$ implies $z_n = z_m$.

Note that (Q5) implies the variability collapse, that is all the within-class representations collapse to their class means. When this condition holds and recall the definition of balanced contrastive loss, for an instance x_i with label y in a batch B, balanced contrastive loss has the following expression:

$$\begin{aligned} \mathcal{L}_{BCL} &= \sum_{B \in \mathcal{B}} \sum_{y \in \mathcal{Y}} \sum_{i \in B_y} \mathcal{L}_i \\ \mathcal{L}_i &= -\log \frac{\exp(z_i \cdot z_{c_y})}{\exp(z_i \cdot z_{c_y}) + \sum_{j \in \mathcal{Y} \setminus \{y\}} \exp(z_i \cdot z_{c_j})} \end{aligned} \tag{13}$$

Note that under the condition of (Q5), for every $B \in \mathcal{B}$, every $y \in \mathcal{Y}$ and every $i \in B_y$, it holds that $z_i = z_{c_y}$, and the label configuration of $\mathcal{L}_i$ is balanced. To minimize the above loss, the solution obviously satisfies the simplex configuration. Leveraging the lower bound of supervised contrastive loss under balanced settings [2], we have

$$\mathcal{L}_i \overset{(Q6)}{\geq} \log\left(1 + (K-1)\exp\left(-\frac{K}{K-1}\right)\right) \tag{14}$$

(Q6) $z_{c_1}, \ldots, z_{c_K}$ *form a regular simplex*

Combine the aforementioned conditions, we can obtain the claimed lower bound of balanced contrastive loss:

$$\mathcal{L}_{BCL}(Z;Y) \geq |\mathcal{D}| \log \left(1 + (K-1)\exp\left(-\frac{K}{K-1}\right)\right) \tag{15}$$

Recall that Z is an N point configuration with labels Y, the equality of Eq. 15 is attained if and only if the following conditions hold. There are $\zeta_1, \ldots, \zeta_K \in \mathbb{R}^h$ such that:

(1) $\forall n \in [N] : z_n = \zeta_{y_n}$

(2) $\zeta_1, \ldots, \zeta_K$ *form a regular simplex*

B. Gradient Analysis

Balanced contrastive loss achieves the balance by averaging the parts of each class. An analysis of the gradients well reflects this conclusion. First, we will discus the defects of the supervised contrastive loss when working on the long-tailed data. Next, we will give the gradient derivation of the balanced contrastive loss, from which we can easily identify that balanced contrastive loss is better at handling long-tailed data.

Recall the definitions of supervised contrastive (SC) loss, neglecting the hyper parameter temperature τ, the gradient of SC loss has the following formulation [3]:

$$\frac{\partial \mathcal{L}_i^{SC}}{\partial z_i} = \underbrace{\sum_{p \in B_y \setminus \{i\}} z_p \left(P_{ip} - \frac{1}{|B_y| - 1}\right)}_{positive\ term} + \underbrace{\sum_{n \in B_y^C} z_n P_{in}}_{negative\ term} \tag{16}$$

where B_y^C is the complement set of B_y and we have defined:

$$\begin{aligned} P_{ip} &= \frac{\exp(z_i \cdot z_p)}{\sum_{k \in B \setminus \{i\}} \exp(z_i \cdot z_k)} \\ P_{in} &= \frac{\exp(z_i \cdot z_n)}{\sum_{k \in B \setminus \{i\}} \exp(z_i \cdot z_k)} \end{aligned} \tag{17}$$

Since there is a normalization function before computing the loss. Let w_i denote the output prior to normalization in a slight abuse of notation, i.e., $z_i = w_i / \|w_i\|$. Then, the gradient with respect to w_i is as follows:

$$\begin{aligned} \frac{\partial \mathcal{L}_i^{SC}}{\partial w_i} &= \frac{1}{\|w_i\|}(I - z_i z_i^T) \left(\sum_{p \in B_y \setminus \{i\}} z_p \left(P_{ip} - \frac{1}{|B_y| - 1}\right) + \sum_{n \in B_y^C} z_n P_{in} \right) \\ &= \frac{1}{\|w_i\|} \left(\underbrace{\sum_{p \in B_y \setminus \{i\}} (z_p - (z_i \cdot z_p) z_i)(P_{ip} - \frac{1}{|B_y| - 1})}_{positive\ term} + \underbrace{\sum_{n \in B_y^C} (z_n - (z_i \cdot z_n) z_i) P_{in}}_{negative\ term} \right) \end{aligned} \tag{18}$$

We mainly concern with the gradients from the negative term. For hard negatives, $z_i \cdot z_n \approx 0$ (assume $z_i \cdot z_n \leq 0$), so that the gradient of $\mathcal{L}_i^{SC}$ from the hard negatives is as follows:

$$\begin{aligned} &\sum_{n \in B_y^C} \|z_n - (z_i \cdot z_n) z_i\| |P_{in}| \\ &\approx \sum_{n \in B_y^C} |P_{in}| \\ &= \sum_{n \in B_y^C} \frac{1}{\sum_{k \in B \setminus \{i\}} \exp(z_i \cdot z_k)} \end{aligned} \tag{19}$$

Given an anchor, the term in the denominator is consistent for all negative samples, resulting in the negative class gradient is proportional to the number of samples. But under the long-tailed distribution, within almost every mini-batch, there are much more head class samples than tail class samples. This leads to all classes being as far away from the head category as possible, and results in an unbalanced feature space.

For balanced contrastive (BC) loss, the gradient has the following formulation:

$$\begin{aligned}\frac{\partial \mathcal{L}_i^{BC}}{\partial z_i} &= -\frac{1}{|B_y|-1}\sum_{p\in B_y\setminus\{i\}}\left(z_p - \sum_{j\in\mathcal{Y}}\frac{1}{|B_j|}\sum_{k\in B_j} z_k X_{ik}\right) \\ &= -\frac{1}{|B_y|-1}\sum_{p\in B_y\setminus\{i\}}\left(z_p - \frac{1}{|B_y|-1}\sum_{p'\in B_y\setminus\{i\}} z_{p'}X_{ip'} - \sum_{j\in\mathcal{Y}\setminus\{y\}}\frac{1}{|B_j|}\sum_{k\in B_j} z_k X_{ik}\right) \\ &= \underbrace{\frac{1}{|B_y|-1}\sum_{p\in B_y\setminus\{i\}} z_p(X_{ip}-1)}_{positive\ term} + \underbrace{\sum_{j\in\mathcal{Y}\setminus\{y\}}\frac{1}{|B_j|}\sum_{k\in B_j} z_k X_{ik}}_{negative\ term}\end{aligned} \tag{20}$$

where we have defined:

$$\begin{aligned}X_{ip} &= \frac{\exp(z_i\cdot z_p)}{\sum_{j\in\mathcal{Y}}\frac{1}{|B_j|}\sum_{k\in B_j}\exp(z_i\cdot z_k)} \\ X_{ik} &= \frac{\exp(z_i\cdot z_k)}{\sum_{j\in\mathcal{Y}}\frac{1}{|B_j|}\sum_{k\in B_j}\exp(z_i\cdot z_k)}\end{aligned} \tag{21}$$

Similar to the derivation of supervised contrastive loss, the gradient with respect to w_i of balanced contrastive loss is as follows:

$$\frac{\partial \mathcal{L}_i^{BC}}{\partial w_i} = \frac{1}{\|w_i\|}\left(\underbrace{\frac{1}{|B_y|-1}\sum_{p\in B_y\setminus\{i\}}(z_p-(z_i\cdot z_p)z_i)(X_{ip}-1)}_{positive\ term} + \underbrace{\sum_{j\in\mathcal{Y}\setminus\{y\}}\frac{1}{|B_j|}\sum_{k\in B_j}(z_k-(z_i\cdot z_k)z_i)X_{ik}}_{negative\ term}\right) \tag{22}$$

Intuitively, balanced contrastive loss balances the gradients from negative classes, avoiding a tremendous gradient update from the negative head class samples. It retains several good properties of supervised contrastive loss. Easy negatives $z_i\cdot z_k \approx -1$ contributes less gradient while hard negatives more gradient, and easy positives $z_i\cdot z_p \approx 1$ (assume $z_i\cdot z_p \geq 0$), contributes less gradient compared with hard positives. In addition to these common properties, the balanced contrastive loss is better at feature alignment, where points belonging to the same class are pulled together. Since almost every mini-batch is long-tailed, for these head class anchors, the gradients in Eq. 18 from the positives will be much larger than when the anchor is tails. It results in tail class samples being unconcerned to pulling these points together. Comparing Eq. 18 with Eq. 22, balanced contrastive loss also adjusts the gradients from the positives, eliminating excessive gradient fluctuations caused by having different anchor classes in different batches and allowing the points of tail classes been pulled closer.

C. More Results

C.1. Ablations of Different Forms of Prototypes.

We compare our method with the other two implementations of the prototype. The first one is using the exponential moving average to calculate the prototype. The second one is using learnable parameters [1, 5]. As shown in Table 1, our implementation achieves the best results and the other two implementations achieve similar results.

C.2. Ablations of Different Configurations of Views.

We compare our configuration with the other two configurations of views. We use the simple augmentation method, i.e., SimAug, to generate both views for contrastive learning. We further have one of the views generated via a stronger augmentation method, i.e., RandAug, or both of the views generated by RandAug. As shown in Table 2, stronger argumentation yields better performance.

Methods	Many	Medium	Few	All
Exponential Moving Average	69.7	54.4	31.9	53.0
Learnable Parameters	68.9	54.0	34.3	53.3
Ours	69.7	53.8	35.5	53.9

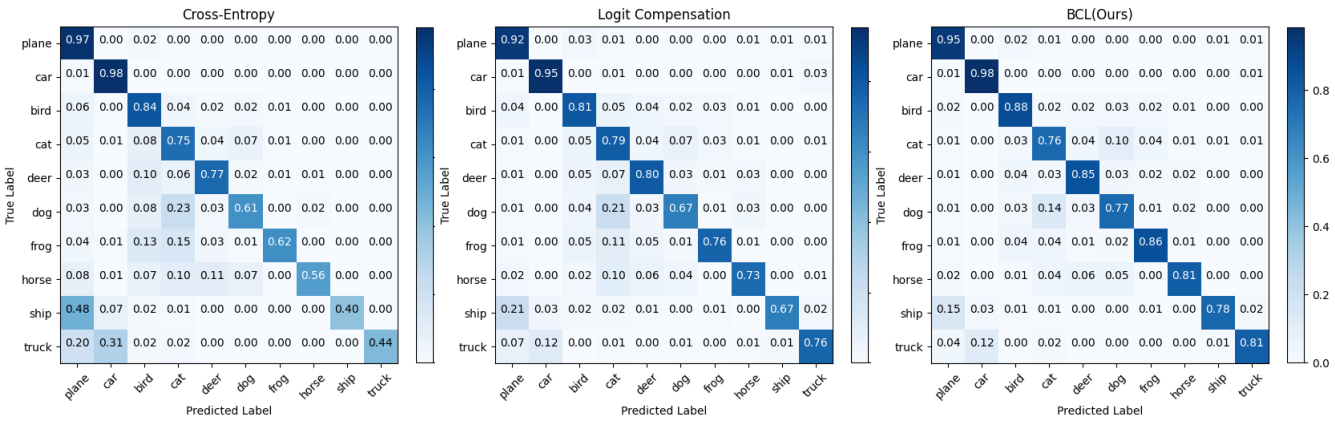
Table 1. Ablation study for different implementations of prototypes on CIFAR-100-LT with an imbalance fator of 100. All models run for 400 epochs with the same training scheme.

Methods	Many	Medium	Few	All
SimAug.&SimAug.	67.2	53.9	36.5	56.7
RandAug.&SimAug.	67.1	54.6	37.1	57.1
RandAug.&RandAug.	67.6	54.6	37.5	57.3

Table 2. Ablation study for different configurations of views on ImageNet-LT. All models run for 90 epochs with the same training scheme.

C.3. Confusion Matrix.

To clearly show where the models are getting confused on long-tailed data, we illustrate the confusion matrix of prediction results on CIFAR-10-LT in Figure 1. With vanilla cross-entropy, the model tends to misclassify low-frequency artifactory categories as high-frequency artifactory categories and low-frequency animal classes as high-frequency animal classes. With logit compensation, misclassification of low-frequency classes is greatly eased. With the proposed BCL, low-frequency classes are more correctly predicted than high-frequency classes, and the accuracies of high-frequency classes are also improved.



(a) (b) (c)

Figure 1. Illustration of confusion matrix of prediction results on CIFAR-10-LT for different models.

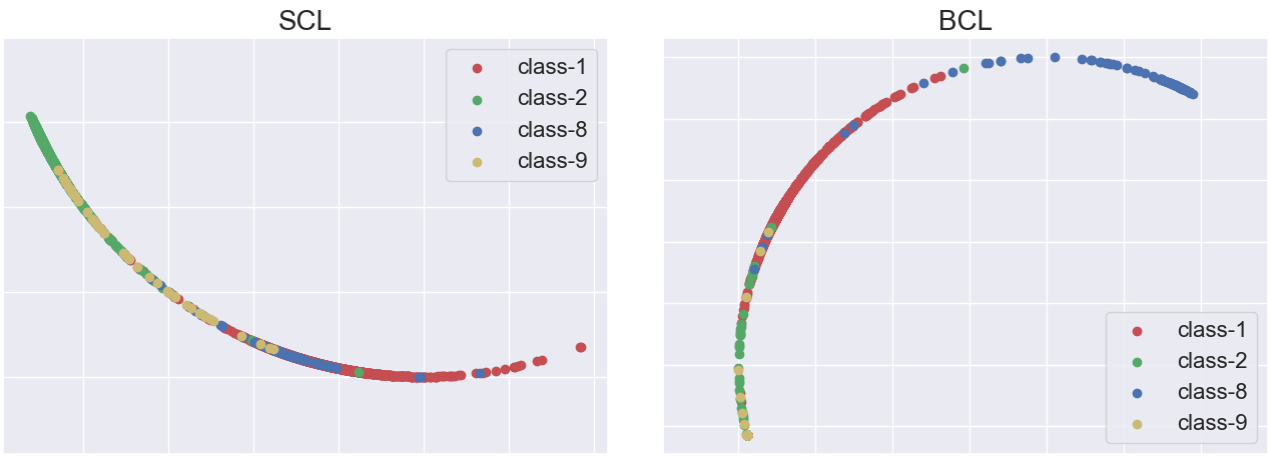
C.4. Visualization of Learned Features.

Similar to [4], we visualize the 2-dimensional MLP output feature learned by SCL and BCL on CIFAR-10-LT. Features of different classes learned by BCL distribute more uniform on the sphere and are more separable than SCL.

References

[1] 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. 5

[2] Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised constrastive learning. In *International Conference on Machine Learning*, pages 3821–3830. PMLR, 2021. 1, 3



(a) (b)

Figure 2. Illustration of features learned by SCL and BCL.

[3] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33, 2020. 4

[4] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. *arXiv preprint arXiv:2111.13998*, 2021. 6

[5] Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. Contrastive learning based hybrid networks for long-tailed image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 943–952, 2021. 5