Supplementary Material of
Decoupled Contrastive Learning for Long-tailed Recognition

Anonymous submission

This supplementary material provides proof to the analysis in Section 3.1 and Section 3.2, and more details of analysis on each proposed component.

Proof to the analysis in Section 3.1

$$
\begin{aligned}
\mathcal{L}_{scl} &= \frac{-1}{|P_i|+1} \sum_{\mathrm{z}_t \in \{\mathrm{z}_i^+ \cup P_i\}} \log \frac{\exp(\mathrm{z}_t \cdot \mathrm{z}_i/\tau)}{\sum\limits_{\mathrm{z}_m \in \{\mathrm{z}_i^+ \cup M\}} \exp(\mathrm{z}_m \cdot \mathrm{z}_i/\tau)}, \\
\frac{\partial \mathcal{L}_{scl}}{\partial \mathrm{z}_i} &= \frac{-1}{|P_i|+1} \sum_{\mathrm{z}_t \in \{\mathrm{z}_i^+ \cup M\}} \frac{\partial}{\partial \mathrm{z}_i} \left\{ \frac{\mathrm{z}_t \cdot \mathrm{z}_i}{\tau} - \log \left(\exp(\mathrm{z}_i^+ \cdot \mathrm{z}_i/\tau) + \sum_{\mathrm{z}_m \in M} \exp(\mathrm{z}_m \cdot \mathrm{z}_i/\tau) \right) \right\}, \\
&= \frac{-1}{\tau(|P_i|+1)} \sum_{\mathrm{z}_t \in \{\mathrm{z}_i^+ \cup M\}} \left\{ \mathrm{z}_t - \frac{\mathrm{z}_i^+ \exp(\mathrm{z}_i^+ \cdot \mathrm{z}_i/\tau) + \sum_{\mathrm{z}_m \in M} \mathrm{z}_m \exp(\mathrm{z}_m \cdot \mathrm{z}_i/\tau)}{\exp(\mathrm{z}_i^+ \cdot \mathrm{z}_i/\tau) + \sum_{\mathrm{z}_m \in M} \exp(\mathrm{z}_m \cdot \mathrm{z}_i/\tau)} \right\}, \\
&= \frac{-1}{\tau(|P_i|+1)} \sum_{\mathrm{z}_t \in \{\mathrm{z}_i^+ \cup M\}} \left\{ \mathrm{z}_t - \sum_{\mathrm{z}_t^* \in \{\mathrm{z}_i^+ \cup M\}} \mathrm{z}_t^* p(\mathrm{z}_t^* | \mathrm{z}_i) - \sum_{\mathrm{z}_n \in N(i)} \mathrm{z}_n p(\mathrm{z}_n | \mathrm{z}_i) \right\}, \\
&= \frac{-1}{\tau(|P_i|+1)} \left\{ \sum_{\mathrm{z}_t \in \{\mathrm{z}_i^+ \cup M\}} \mathrm{z}_t - \sum_{\mathrm{z}_t \in \{\mathrm{z}_i^+ \cup M\}} (|P_i|+1) \mathrm{z}_t p(\mathrm{z}_t | \mathrm{z}_i) - \sum_{\mathrm{z}_n \in N_i} (|P_i|+1) \mathrm{z}_n p(\mathrm{z}_n | \mathrm{z}_i) \right\}, \\
&= \frac{1}{\tau} \left\{ \sum_{\mathrm{z}_n \in N_i} \mathrm{z}_n p(\mathrm{z}_n | \mathrm{z}_i) + \mathrm{z}_i^+ \left(p(\mathrm{z}_i^+ | \mathrm{z}_i) - \frac{1}{(|P_i|+1)} \right) + \sum_{\mathrm{z}_t \in P_i} \mathrm{z}_t \left(p(\mathrm{z}_t | \mathrm{z}_i) - \frac{1}{(|P_i|+1)} \right) \right\}.
\end{aligned}
\tag{1}
$$

At the beginning of the training, the model is randomly initialized, we can assume $p(\mathrm{z}_m|\mathrm{z}_i) \approx 1/|M|, \mathrm{z}_m \in \{\mathrm{z}_i^+ \cup M\}$,

$$
\begin{aligned}
\left. \frac{\partial \mathcal{L}_{scl}}{\partial \mathrm{z}_i} \right|_{\mathrm{z}_i^+} &= \mathrm{z}_i^+ \left(p(\mathrm{z}_i^+ \mid \mathrm{z}_i) - \frac{1}{|P_i|+1} \right), \\
\left. \frac{\partial \mathcal{L}_{scl}}{\partial \mathrm{z}_i} \right|_{\mathrm{z}_t} &= \mathrm{z}_t \left(p(\mathrm{z}_t | \mathrm{z}_i) - \frac{1}{|P_i|+1} \right), \mathrm{z}_t \in P_i,
\end{aligned}
\tag{2}
$$

since the feature is normalized, $\|\mathrm{z}_i\|_2 = 1$,

$$
\begin{aligned}
\left\| \left. \frac{\partial \mathcal{L}_{scl}}{\partial \mathrm{z}_i} \right|_{\mathrm{z}_t} \right\|_2 &\approx \left| \frac{1}{|M|} - \frac{1}{|P_i|+1} \right|, \\
\sum_{\mathrm{z}_t \in P_i} \left\| \left. \frac{\partial \mathcal{L}_{scl}}{\partial \mathrm{z}_i} \right|_{\mathrm{z}_t} \right\|_2 &\approx |P_i| \left| \frac{1}{|M|} - \frac{1}{|P_i|+1} \right|, \\
\frac{\left\| \left. \frac{\partial \mathcal{L}_{scl}}{\partial \mathrm{z}_i} \right|_{\mathrm{z}_i^+} \right\|_2}{\sum\limits_{\mathrm{z}_t \in P_i} \left\| \left. \frac{\partial \mathcal{L}_{scl}}{\partial \mathrm{z}_i} \right|_{\mathrm{z}_t} \right\|_2} &\approx \frac{1}{|P_i|}.
\end{aligned}
\tag{3}
$$

When SCL converges,

$$\begin{aligned} \frac{\partial \mathcal{L}_{scl}}{\partial \mathrm{z}_i} &= 0, \\ p(\mathrm{z}_i^+|\mathrm{z}_i) &= \frac{1}{|P_i|+1}. \end{aligned} \tag{4}$$

Proof to the analysis in Section 3.2

Similar to the proof to the analysis in Section 3.1,

$$\begin{aligned} \mathcal{L}_{dscl} &= \frac{-1}{|P_i|+1} \sum_{\mathrm{z}_t \in \{\mathrm{z}_i^+ \cup P_i\}} \log \frac{\exp w_t(\mathrm{z}_t \cdot \mathrm{z}_i/\tau)}{\sum\limits_{\mathrm{z}_m \in \{\mathrm{z}_i^+ \cup M\}} \exp(\mathrm{z}_m \cdot \mathrm{z}_i/\tau)}, \\ \frac{\partial \mathcal{L}_{dscl}}{\partial \mathrm{z}_i} &= \frac{-1}{|P_i|+1} \sum_{\mathrm{z}_t \in \{\mathrm{z}_i^+ \cup M\}} \frac{\partial}{\partial \mathrm{z}_i} \left\{ \frac{w_t(\mathrm{z}_t \cdot \mathrm{z}_i)}{\tau} - \log \left(\exp(\mathrm{z}_i^+ \cdot \mathrm{z}_i/\tau) + \sum_{\mathrm{z}_m \in M} \exp(\mathrm{z}_m \cdot \mathrm{z}_i/\tau) \right) \right\}, \\ &= \frac{-1}{\tau(|P_i|+1)} \sum_{\mathrm{z}_t \in \{\mathrm{z}_i^+ \cup M\}} \left\{ w_t \mathrm{z}_t - \frac{\mathrm{z}_i^+ \exp(\mathrm{z}_i^+ \cdot \mathrm{z}_i/\tau) + \sum_{\mathrm{z}_m \in M} \mathrm{z}_m \exp(\mathrm{z}_m \cdot \mathrm{z}_i/\tau)}{\exp(\mathrm{z}_i^+ \cdot \mathrm{z}_i/\tau) + \sum_{\mathrm{z}_m \in M} \exp(\mathrm{z}_m \cdot \mathrm{z}_i/\tau)} \right\}, \\ &= \frac{-1}{\tau(|P_i|+1)} \sum_{\mathrm{z}_t \in \{\mathrm{z}_i^+ \cup M\}} \left\{ w_t \mathrm{z}_t - \sum_{\mathrm{z}_t^* \in \{\mathrm{z}_i^+ \cup M\}} \mathrm{z}_t^* p(\mathrm{z}_t^*|\mathrm{z}_i) - \sum_{\mathrm{z}_n \in N(i)} \mathrm{z}_n p(\mathrm{z}_n|\mathrm{z}_i) \right\}, \\ &= \frac{-1}{\tau(|P_i|+1)} \left\{ \sum_{\mathrm{z}_t \in \{\mathrm{z}_i^+ \cup M\}} w_t \mathrm{z}_t - \sum_{\mathrm{z}_t \in \{\mathrm{z}_i^+ \cup M\}} (|P_i|+1) \mathrm{z}_t p(\mathrm{z}_t|\mathrm{z}_i) - \sum_{\mathrm{z}_n \in N_i} (|P_i|+1) \mathrm{z}_n p(\mathrm{z}_n|\mathrm{z}_i) \right\}, \\ &= \frac{1}{\tau} \left\{ \sum_{\mathrm{z}_n \in N(i)} \mathrm{z}_n p(\mathrm{z}_n|\mathrm{z}_i) + \mathrm{z}_i^+ \left(p(\mathrm{z}_i^+|\mathrm{z}_i) - \frac{\alpha(|P_i|+1)}{(|P_i|+1)} \right) + \sum_{\mathrm{z}_t \in P_i} \mathrm{z}_t \left(p(\mathrm{z}_t|\mathrm{z}_i) - \frac{(1-\alpha)(|P_i|+1)}{(|P_i|+1)|P_i|} \right) \right\}, \\ &= \frac{1}{\tau} \left\{ \sum_{\mathrm{z}_n \in N_i} \mathrm{z}_n p(\mathrm{z}_n|\mathrm{z}_i) + \mathrm{z}_i^+ \left(p(\mathrm{z}_i^+|\mathrm{z}_i) - \alpha \right) + \sum_{\mathrm{z}_t \in P_i} \mathrm{z}_t \left(p(\mathrm{z}_t|\mathrm{z}_i) - \frac{1-\alpha}{|P_i|} \right) \right\}. \end{aligned} \tag{5}$$

Similarly, assume $p(\mathrm{z}_m|\mathrm{z}_i) \approx 1/|M|, \mathrm{z}_m \in \{\mathrm{z}_i^+ \cup M\}$,

$$\begin{aligned} \left.\frac{\partial \mathcal{L}_{dscl}}{\partial \mathrm{z}_i}\right|_{\mathrm{z}_i^+} &= \mathrm{z}_i^+ \left(p(\mathrm{z}_i^+|\mathrm{z}_i) - \alpha \right), \\ \left.\frac{\partial \mathcal{L}_{dscl}}{\partial \mathrm{z}_i}\right|_{\mathrm{z}_t} &= \mathrm{z}_t \left(p(\mathrm{z}_t|\mathrm{z}_i) - \frac{1-\alpha}{|P_i|} \right), \mathrm{z}_t \in P_i, \\ \left\| \left.\frac{\partial \mathcal{L}_{dscl}}{\partial \mathrm{z}_i}\right|_{\mathrm{z}_i^+} \right\|_2 &\approx \left| \frac{1}{|M|} - \alpha \right|, \\ \sum_{\mathrm{z}_t \in P_i} \left\| \left.\frac{\partial \mathcal{L}_{dscl}}{\partial \mathrm{z}_i}\right|_{\mathrm{z}_t} \right\|_2 &\approx |P_i| \left| \frac{1}{|M|} - \frac{1-\alpha}{|P_i|} \right|, \end{aligned} \tag{6}$$

$|M| \gg 1$. Therefore, we can assume $\frac{1}{|M|} \approx 0$, thus,

$$\frac{\left\| \left.\frac{\partial \mathcal{L}_{dscl}}{\partial \mathrm{z}_i}\right|_{\mathrm{z}_i^+} \right\|_2}{\sum\limits_{\mathrm{z}_t \in P_i} \left\| \left.\frac{\partial \mathcal{L}_{dscl}}{\partial \mathrm{z}_i}\right|_{\mathrm{z}_t} \right\|_2} \approx \frac{\alpha}{1-\alpha}. \tag{7}$$

When DSCL converges,

$$\begin{aligned} \frac{\partial \mathcal{L}_{dscl}}{\partial \mathrm{z}_i} &= 0 \\ p(\mathrm{z}_i^+|\mathrm{z}_i) &= \alpha. \end{aligned} \tag{8}$$

More Experimental Results on Each Proposed Component

Ablation Study on Datasets with Different Imbalanced Ratios We conduct more experiments to further validate the effectiveness of the proposed method across different imbalanced ratios. We generate 3 datasets with different imbalanced ratios from the ImageNet1K following the Pareto distribution. The detailed statistics of the generated datasets and the experimental results are shown in Table 1 and Table 2, respectively. As shown in the Table 2, our proposed DSCL and PBSD generalize well on different imbalanced ratios. Both of them can bring performance improvement. Some long-tailed methods may be harmful for the performance on a balanced dataset. The experimental result on a balanced Dataset C also show that our DSCL does not decrease the accuracy on a balanced dataset and PBSD can also bring performance improvement. This result further validate the generalization ability of our method on a balanced dataset.

	Max	Min	Total	Imbalanced Ratio
ImageNet-LT	1280	5	115846	256
Dataset A	857	10	115852	85.7
Dataset B	343	37	115801	9.27
Dataset C	115	115	115000	1

Table 1: Dataset statistics on training instance numbers, including the maximal and minimal instance number per class, the number of total training instances, and the imbalanced ratio.

	ImageNet-LT	Dataset A	Dataset B	Dataset C
Baseline	51.2	53.2	56.1	60.5
DSCL	52.6	54.7	57.3	60.7
PBSD	56.3	59.2	61.8	65.1
DSCL + PBSD	57.7	59.6	62.7	65.2

Table 2: Ablation study of each component in our method on datasets with different imbalanced ratios. SCL (Khosla et al. 2020) is used as baseline.

Ablation study of DSCL. We conduct experiments to validate that DSCL is a reasonable choice to remove the bias of SCL. As discussed previous Section, the bias of SCL is caused by the imbalanced M. Therefore, one possible solution is to maintain a balanced memory queue M. The result is shown in the second row of Table 3. The balanced memory queue does not bring performance improvement and is even harmful to the performance of tail classes. The reason is that the balanced memory queue leads to more negative gradients to the tail classes, *i.e.*, the tail classes receive too many gradients to push them away from other samples in the feature space. Re-weighting is also a commonly used method to remove the bias. Based on SCL, we add the loss weight on each instance as in (Cui et al. 2019). The result shows that re-weighting method decreases the performance. The same phenomenon is also validated in (Kang et al. 2019) that the re-weighting methods lead to a poor discriminative feature space. It can be concluded that DSCL is a reasonable and non-trivial solution to remove the bias of SCL.

Ablation Study of Different Scale Ranges The patch-based features are used to extract visual patterns. The scale range $[s_1, s_2]$ controls the size of the area to get patch-based features. We further conduct experiments to validate the influence of this parameter on the performance. The results are summarized at Table 4. It is important to set a small s_1, *e.g.*, $s_1 = 0.05$. Keeping $s_2 = 0.6$ and increasing s_1 from 0.05 to 0.3 decrease the accuracy by about 1%. This result is consistent with our motivation, *i.e.*, using a small part of an object extracts visual pattern. Setting the scale range as $[1.0, 1.0]$ degenerates the patch-based features to global features, decreasing the accuracy from 57.7% to 56.2%. The importance of introducing patch-based features in PBSD can be further validated.

Class activation map visualization is shown in Fig. 1. The CAM of the model without PBSD focuses on smaller areas, while PBSD makes the model capture more useful cues. This result can further validate the effectiveness of our method.

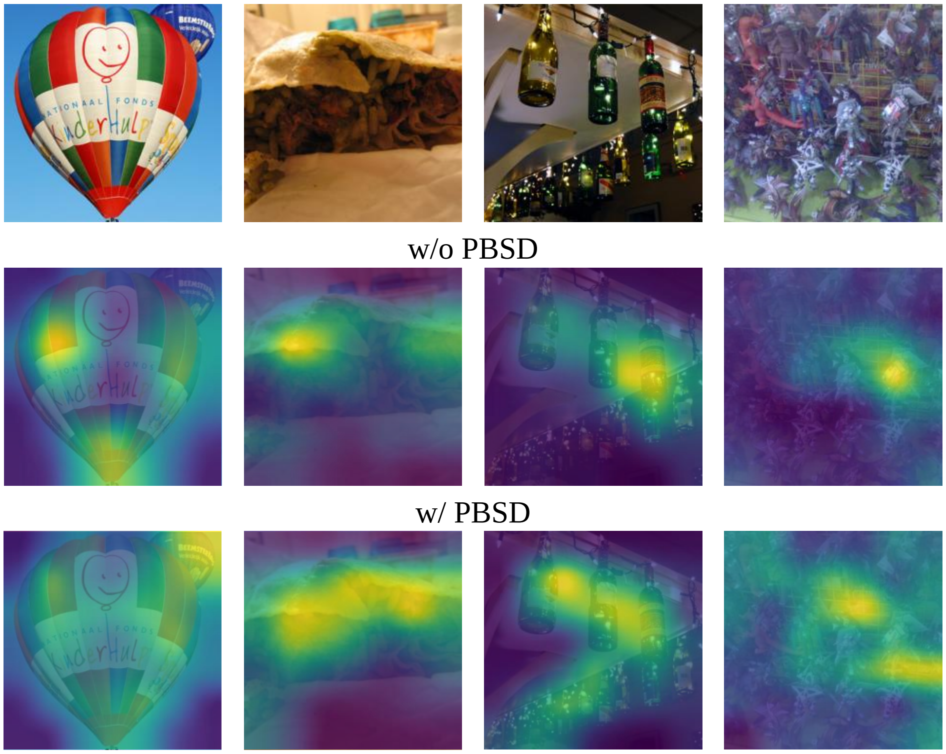
Patch boxes Generation of Patch-based Feature To extract visual patterns, we introduce the patch-based features, which

Settings	Many	Medium	Few	Overall
Baseline	61.6	48.6	30.3	51.2
Balanced Queue	62.3	48.9	29.2	51.4
Re-weighting	59.7	45.9	30.4	49.1
DSCL	**63.4**	**50.0**	**31.4**	**52.6**

Table 3: Ablation study of DSCL on ImageNet-LT. Top-1 accuracy is used as metric. Balanced Queue denotes M is a balanced memory queue. Re-weighting denotes the loss weight is added on each instance as in (Cui et al. 2019).

Scale Ranges	Acc
$[0.05, 0.7]$	57.5
$[0.05, 0.6]$	57.7
$[0.05, 0.5]$	57.2
$[0.1, 0.7]$	57.4
$[0.1, 0.6]$	57.4
$[0.1, 0.5]$	57.2
$[0.15, 0.6]$	57.1
$[0.3, 0.6]$	56.7
$[1.0, 1.0]$	56.2

Table 4: Ablation study of different scale ranges to get patch-based features on ImageNet-LT.



w/o PBSD

w/ PBSD

Figure 1: CAM (Zhou et al. 2016) visualizations of our method with/without PBSD on images sampled from tailed classes of ImageNet-LT.

are pooled with ROI pooling from the feature map of the global view at some randomly generated positions. We here give the pseudo-code of patch boxes generation in Alg. 1. The ratio range is set as $[0.75, 1.33]$.

```
Algorithm 1: Patch boxes Generation of Patch-based Feature
    Input:
    Global view of the image $x_i$ whose size is $h \times w$
    The number of patch boxes $L$
    Scale range $[s_1, s_2]$
    Ratio range $[r_1, r_2]$
    Output:
    Coordinates $\{B_i[j]\}_{j=1}^{L}$ to get patch-based feature

    for $j = 1$ to L do
        $s \sim U(s_1, s_2)$                                ▹ Get scale from uniform distribution
        $r \sim U(r_1, r_2)$                                ▹ Get ratio from uniform distribution
        $h_p = s \cdot r \cdot h$                           ▹ Height of the area
        $w_p = s \cdot w$                                   ▹ Width of the area
        $x_p = (h - h_p) \cdot u, u \sim U(0, 1)$           ▹ Top-left x-coordinate
        $y_p = (w - w_p) \cdot v, v \sim U(0, 1)$           ▹ Top-left y-coordinate
        $B_i[j] = (x_p, y_p, h_p + x_p, w_p + y_p)$
    end for
    Return: $\{B_i[j]\}_{j=1}^{L}$
```

Pseudo-Code of Patch-based Self Distillation We here give the pseudo-code of patch-based self distillation in Alg. 2.

```
Algorithm 2: Patch-based Self Distillation
    Input:
    Global view of the image $x_i$ whose size is $h \times w$
    Backbone Model $f_\theta$
    Projection Head $g_\gamma$
    Memory Bank $M$
    Feature Embedding $z_i^+$ of Another Data Augmentation
    Output:
    PBSD Loss Function $\mathcal{L}_{pbsd}$

    $\{B_i[j]\}_{j=1}^{L} \leftarrow$ Alg. 1$(x_i)$          ▹ Generate coordinates to get patch-based feature with Alg. 1
    $u_i = f_\theta(x_i)$                                     ▹ Get global feature map of the image
    $\{c_i[j]\}_{j=1}^{L} = \{g_\gamma(\mathrm{ROI}(u_i, B_i[j]))\}_{j=1}^{L}$     ▹ Get patch-based features thorough ROI pooling
    $\{s_i[j]\}_{j=1}^{L} = \{g_\gamma(f_\theta(\mathrm{Crop}(x_i, B_i[j])))\}_{j=1}^{L}$     ▹ Crop multi image patches and extract their feature embeddings
    $\mathcal{L}_{pbsd} = \frac{1}{L}\sum_{j=1}^{L}\sum_{z_t \in \{z_i^+ \cup M\}} -p(z_t|c_i[j]) \log p(z_t|s_i[j])$     ▹ Calculate the patch-based self distillation loss
    Return: $\mathcal{L}_{pbsd}$
```

References

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *CVPR*, 9268–9277.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2019. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *CVPR*, 2921–2929.