In this supplementary material, we present an additional study on Feature Clusters Compression (FCC). Appendix A introduces a detailed formula derivation for the feasibility demonstration of FCC (described in Section 3.2 of our paper). Appendix B presents implementation details that are not described in Sections 4.1 and 4.2. An evaluation of different compression strategies on different datasets is presented in Appendix C. Pseudo-code for FCC and ECE analysis can be found in Appendix D and Appendix E, respectively. Visualization of features is shown in Appendix F.

A. Detailed Formula Derivation

Detailed formula derivation for the feasibility demonstration of FCC is provided in this part.

Setting. We present a Fully Connected (FC) network of binary classification to prove the feasibility, and its architecture is shown in Fig. 1. The FC network contains an input layer with 3 neurons, a hidden layer with 3 neurons $\{a_1, a_2, a_3\}$ and an output layer with 2 neurons $\{o_1, o_2\}$. $\{\tau x_1, \tau x_2, \tau x_3\}$ and $\{x_1, x_2, x_3\}$ are the multiplied and original features, respectively, and they both belong to class 1. The scaling factor of class 1 is τ ($\tau > 1$). $\{y_1, y_2\}$ and $\{y_1{}', y_2{}'\}$ are outputs of the multiplied and original feature produced by the FC network, respectively. $\{w_{i1}, w_{i2}, w_{i3}\}$ and b_i are weights and bias of the neuron a_i ($i \in \{1, 2, 3\}$), respectively. $\{n_{j1}, n_{j2}, n_{j3}\}$ and z_j are weights and bias of the neuron o_j ($j \in \{1, 2\}$), respectively.

If the FC network can normally work, the classification result of the original feature will be equal to that of the multiplied feature, i.e., $y_1{}' > y_2{}'$ when $y_1 > y_2$.

Outputs of the Multiplied and Original Features. The outputs $\{a_1, a_2, a_3\}$ of the hidden layer can be formulated as follows:

$$a_1 = \tau w_{11}x_1 + \tau w_{12}x_2 + \tau w_{13}x_3 + b_1 \quad (1)$$

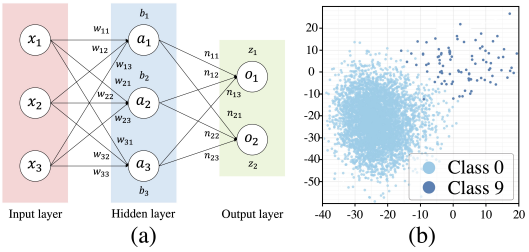
$$a_2 = \tau w_{21}x_1 + \tau w_{22}x_2 + \tau w_{23}x_3 + b_2 \quad (2)$$

$$a_3 = \tau w_{31}x_1 + \tau w_{32}x_2 + \tau w_{33}x_3 + b_3 \quad (3)$$

and the outputs (y_1 and y_2) of the multiplied feature can be expressed as follows:

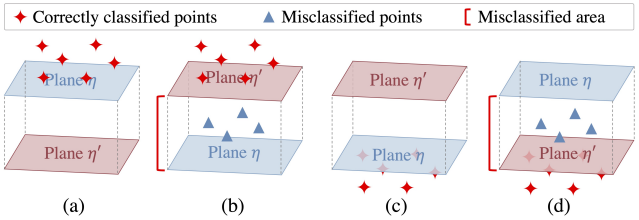
$$\begin{aligned} y_1 &= n_{11}a_1 + n_{12}a_2 + n_{13}a_3 + z_1 \\ &= n_{11}(\tau w_{11}x_1 + \tau w_{12}x_2 + \tau w_{13}x_3) + n_{11}b_1 + \\ &\quad n_{12}(\tau w_{21}x_1 + \tau w_{22}x_2 + \tau w_{23}x_3) + n_{12}b_2 + \\ &\quad n_{13}(\tau w_{31}x_1 + \tau w_{32}x_2 + \tau w_{33}x_3) + n_{13}b_3 + \\ &\quad z_1 \end{aligned} \quad (4)$$

$$\begin{aligned} y_2 &= n_{21}a_1 + n_{22}a_2 + n_{23}a_3 + z_2 \\ &= n_{21}(\tau w_{11}x_1 + \tau w_{12}x_2 + \tau w_{13}x_3) + n_{21}b_1 + \\ &\quad n_{22}(\tau w_{21}x_1 + \tau w_{22}x_2 + \tau w_{23}x_3) + n_{22}b_2 + \\ &\quad n_{23}(\tau w_{31}x_1 + \tau w_{32}x_2 + \tau w_{33}x_3) + n_{23}b_3 + \\ &\quad z_2 \end{aligned} \quad (5)$$



(a) (b)

Figure 1. (a) Architecture of the FC network. (b) Visualization of features of classes 0 and 9 of CIFAR-10-LT-100.



(a) (b) (c) (d)

Figure 2. Relationship between planes η and η' and feature points in geometric space. When feature points are above plane η, (a) plane η' is below plane η, or (b) plane η' is above plane η. When feature points are below plane η, (c) plane η' is above plane η, or (d) plane η' is below plane η.

then we denote $(y_1 - y_2)$ as η, $(w_{11}x_1 + w_{12}x_2 + w_{13}x_3)$ as X_1, $(w_{21}x_1 + w_{22}x_2 + w_{23}x_3)$ as X_2, $(w_{31}x_1 + w_{32}x_2 + w_{33}x_3)$ as X_3 and $(n_{11}b_1 + n_{12}b_2 + n_{13}b_3 + z_1) - (n_{21}b_1 + n_{22}b_2 + n_{23}b_3 + z_2)$ as B, further η is converted as follows:

$$\eta = \tau k_1 X_1 + \tau k_2 X_2 + \tau k_3 X_3 + B \quad (6)$$

where k_i is $(n_{1i} - n_{2i}), i \in \{1, 2, 3\}$. We can notice that η is a (decision) plane in geometric space when $\eta = 0$. Due to $y_1 > y_2$, $\eta > 0$ and Eq. (6) can be formulated as follows:

$$\begin{cases} \tau d_1 X_1 + \tau d_2 X_2 + \tau d_3 X_3 > 1, & B < 0 \\ \tau d_1 X_1 + \tau d_2 X_2 + \tau d_3 X_3 < 1, & B > 0 \end{cases} \quad (7)$$

where d_i denotes $-k_i/B, i \in \{1, 2, 3\}$. In geometric space, the point (X_1, X_2, X_3) is above the plane when $B < 0$, while it is below the plane when $B > 0$. The case of $B = 0$ will be discussed later. On the same principle, we make $y_1' - y_2'$ equal to η', which can be formulated as follows:

$$\eta' = k_1 X_1 + k_2 X_2 + k_3 X_3 + B \quad (8)$$

When $\eta' = 0$, Eq. (8) can be formulated as follows:

$$d_1 X_1 + d_2 X_2 + d_3 X_3 = 1 \quad (9)$$

where η' is also a plane when $\eta' = 0$. We can observe that $\{1/d_1, 1/d_2, 1/d_3\}$ and $\{1/\tau d_1, 1/\tau d_2, 1/\tau d_3\}$ are intercepts of planes η' and η, respectively. And the intercepts of plane η' are τ times of that of plane η, so planes η and η' are parallel in geometric space. Meanwhile, plane η' is either above plane η or below it depending on the intercepts.

Analysis. In the following, we will discuss the relation-

ship between planes η and η' and feature points in geometric space to explore whether y_1' is also greater than y_2' under $y_1 > y_2$.

(1) When $B < 0$, the point (X_1, X_2, X_3) is above plane η based on Eq. (7). If plane η' is below plane η, the point is also above plane η', as shown in Fig. 2a, so $d_1X_1 + d_2X_2 + d_3X_3 > 1$ in Eq. (9), and then we can get $\eta' > 0$ (i.e., $y_1' > y_2'$) based on Eqs. (8) and (9). That implies the FC can normally work on this point.

If plane η' is above plane η, the point might be above or below plane η', as shown in Fig. 2b. The point can also be correctly classified when it is above plane η' since $d_1X_1 + d_2X_2 + d_3X_3 > 1$, but when it is below plane η', $d_1X_1 + d_2X_2 + d_3X_3 < 1$ and $y_1' < y_2'$, which means the FC will misclassify the point.

(2) When $B > 0$, the point is below plane η. If plane η' is above plane η, the point is also blow plane η' as shown in Fig. 2c, so $d_1X_1 + d_2X_2 + d_3X_3 < 1$, and we can get $\eta' > 0$, i.e., $y_1' > y_2'$. On this condition, the FC can also normally classify this point.

If plane η' is below plane η, the point might be above or below plane η', as shown in Fig. 2d. When the point is above plane η', $d_1X_1 + d_2X_2 + d_3X_3 > 1$ and $y_1' < y_2'$, i.e., the FC cannot correctly classify the point. When the point is below plane η', $d_1X_1 + d_2X_2 + d_3X_3 < 1$ and $y_1' > y_2'$, so the FC can normally work this point.

(3) When $B = 0$, planes η and η' coincide with each other, which will make $y_1' > y_2'$ when $y_1 > y_2$.

In a nutshell, the classifier can normally work on original features, except those falling between planes η and η', i.e., "misclassified area" in Fig. 2.

B. Implementation Details

We provide implementation details that are not described in Sections 4.1 and 4.2.

For CIFAR-LT and iNaturalist 2018, we utilize simple data augmentation by applying random cropping and random horizontal flipping to 32×32 and 224×224 size, respectively. For ImageNet-LT, we employ simple random horizontal flips, color jittering, and take random crops 224×224 size. For two-stage training methods, the balanced fine tuning both starts from 160^{th} epoch. The hyper-parameters of compared methods are shown in Tab. 1.

C. Compression Strategies

We evaluate different compression strategies (described in section 3.1) on CIFAR-10-LT-50, CIFAR-10-LT-100 and CIFAR-100-LT-50. The results are shown in Fig. 3, where the top, middle and bottom rows denote the results on CIFAR-10-LT-50, CIFAR-10-LT-100 and CIFAR-100-LT-50, respectively. Consistent with the conclusions described in Section 4.3, in all datasets, equal difference compression outperforms other strategies. Half (Top 50%) compression exhibits the second best results. Uniform and Half (Bottom 50%) compression achieve poor performance. In some cases, they obtain better results than those of baseline (vanilla ResNet-32), but performance improvement is lower than other strategies.

Family	Method	Hyper-parameters
Re-weighting	Focal loss	$\gamma = 2$
	CB Focal loss	$\beta = 0.9$, $\gamma = 1$
	CBCE	$\beta^{\dagger} = 0.9$, $\beta^{\star} = 0.9999$
	CELS	$\epsilon = 0.1$
	CELAS	Smooth head and tail: 0.4 and 0.1
	LDAM	Scale: 30, max margin: 0.5
	CDT	$\gamma = 0.2$
Mixup	Input Mixup	$\alpha = 1$
	Manifold Mixup	$\alpha = 1$, location: pool
	Remix	$\alpha = 1$, $\kappa = 3$, $\tau = 0.5$
Two-stage training	DiVE	Temperature: 2, power: 0.5, $\alpha = 0.5$
Multi-expert	SADE	The number of experts is 3.
	NCL	The number of experts is 3.

Table 1. Implementation details of compared methods.† and $\star$ denote the β is for CIFAR-100-LT and CIFAR-10-LT, respectively.

Furthermore, we also introduce the compression strategy for $\tau < 1$ (Reverse compression), which can be formulated as follows:

$$\tau_i = 1 - \gamma * i/C \tag{10}$$

where $\gamma \in (0, 1]$ is a scaling hyper-parameter, C is the number of classes, and $i \in [0, C)$ is the index of class. The results are shown in Fig. 3, where reverse compression fails to improve raw methods. This is because the original features are expanded rather than compressed, and the feature points are mapped sparsely, making it easier to cross the boundary during testing.

Algorithm 1 Feature Clusters Compression (FCC)

Input: Original backbone features F and their labels d in each batch, batch size B and number of classes N.
Parameter: The scaling hyper-parameter γ
Output: Multiplied features F'.

```
Setting τ for each class.
for i = 0 to (N − 1) do
    τ_i ← 1 + γ ∗ (1 − i/N).
end for
for j = 0 to (B − 1) do
    F_j' ← F_j ∗ τ_{d_j}
end for
return F'
```

Meanwhile, we visualize the recall and confusion matrix between baseline and FCC, as shown in Fig. 4 and Fig. 5, respectively. The results illustrate FCC effectively improves the performance of minority classes. We observe that FCC might damage the performance of some majority classes,

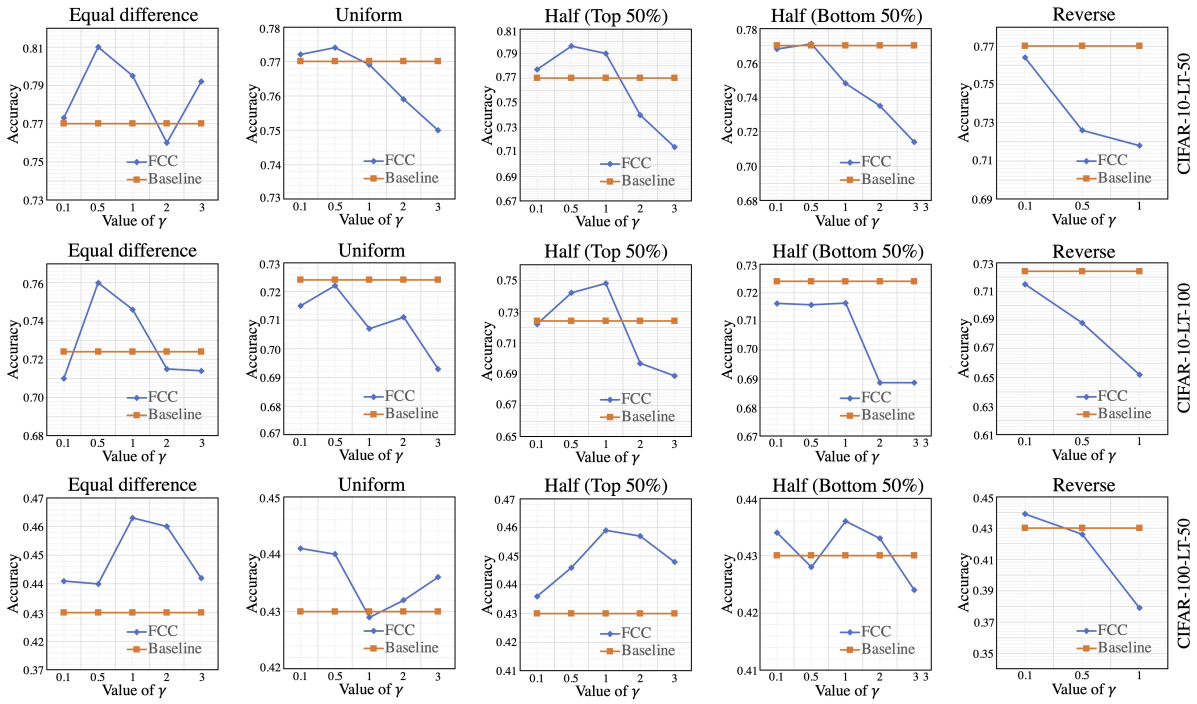


Figure 3. Accuracy comparisons of each compression strategy. The top, middle and bottom rows show the results on CIFAR-10-LT-50, CIFAR-10-LT-100 and CIFAR-100-LT-50, respectively. Baseline is vanilla ResNet-32.

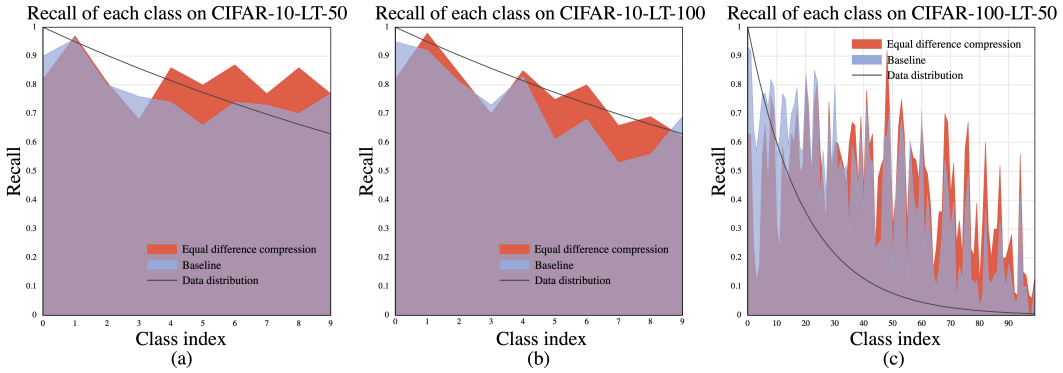


Figure 4. Recall comparisons of each class between baseline (Vanilla ResNet-32) and FCC with equal difference compression (γ is set to 0.5 and 1 on CIFAR-10-LT and CIFAR-100-LT, respectively).

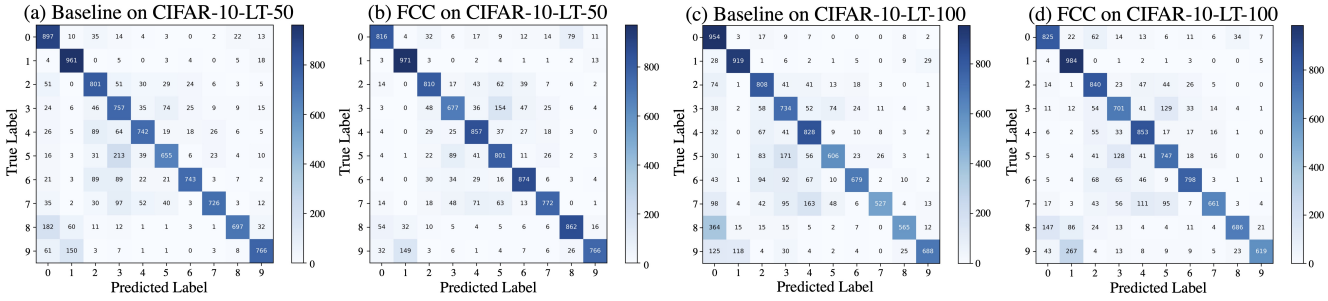


Figure 5. Confusion matrices produced by baseline (vanilla ResNet-32) and FCC on CIFAR-10-LT-50 and CIFAR-10-LT-100.

but it does not affect the overall performance.

D. Pseudo-code of FCC

The pseudo-code of FCC in training procedure is presented in Algorithm 1. FCC can be achieved with a concise code snippet, such that it can be easily applied to any deep neural network.

E. Discuss Expected Calibration Error (ECE).

We show the reliability diagrams of different methods on CIFAR-100-LT-100 in Fig. 6, which illustrate our FCC can not only effectively improve the accuracy (from 39.1% to 41.8%) but also greatly reinforce the network calibration (ECE from 37.8% to 3.46%) for long-tailed recognition.

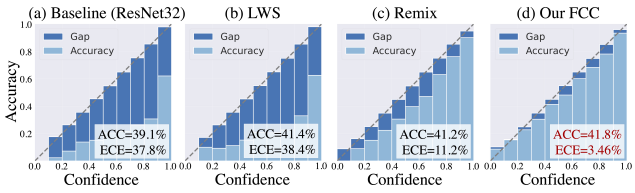


Figure 6. Reliability diagrams of different methods on CIFAR-100-LT-100.

F. Visualization for sparse clusters

In our paper, we describe that the features of minority classes are mapped to sparse clusters relative to those of majority classes. To demonstrate this, we provide the visualization of the features of class 0 and class 9 from CIFAR-10-LT-100, as shown in Fig. 1b, in which class 9 exhibits sparser clusters than class 0.