Appendix A: Implementation Details

In this section, we provide the implementation details of the models used in our experiments. On each dataset, we fix the batch size and training epochs for different baselines based on contrastive loss for a fair comparison.

CIFAR-10-LT and CIFAR-100-LT: We build TSC on top of SimCLR-based supervised contrastive learning on CIFAR-10-LT and CIFAR-100-LT [27], with a ResNet-32 structure [50]. All experiments on CIFAR are performed on 4 NVIDIA Titan X Pascal GPUs. For the first stage, the encoder is trained for 1000 epochs with batch size 1024. The learning rate is set to 0.5 initially and decreased using cosine annealing strategy. For the second stage, the linear classifier is trained for 200 epochs with LDAM loss and class re-weighting with batch size 128. The learning rate is set to 0.1 initially and multiplied by 0.1 at epoch 140, 180 and 190.

ImageNet-LT and iNaturalist: We build TSC on top of MoCo-based supervised contrastive learning on ImageNet-LT and iNaturalist [23], with a ResNet-50 structure [23,51]. All experiments on them are performed on 8 NVIDIA Titan X Pascal GPUs. For the first stage, the encoder is trained for 400 epochs with batch size 256. The learning rate is set to 0.1 initially and decreased using cosine annealing strategy. For the second stage, the linear classifier is trained for 40 epochs with with CE loss and class-balanced sampling with batch size 2048. The learning rate is set to 10 on ImageNet-LT and 30 on iNaturalist initially and multiplied by 0.1 at epoch 20 and 30.

Appendix B: Additional Results

In this section, we provide additional results to better understand TSC.

	CIFAR-10-LT	CIFAR-100-LT
LDAM-DRW [4]	77.1	42.3
KCL [23]	77.6	42.8
TSC	**79.7**	**43.8**

Table 10. Comparison among LDAM-DRW [4], KCL [23] and TSC on CIFAR-10-LT and CIFAR-100-LT for 1000 training epochs.

Epochs	200	400
KCL [23]	51.5	51.6
TSC	**52.3**	**52.4**

Table 11. Comparison between KCL [23] and TSC on ImageNet-LT for 200 and 400 training epochs.

Number of traning epochs: It is standard in the literature to have a longer training time for contrastive learning than non-contrastive learning because it converges more slowly [6, 21, 23, 27]. In Sec. 4, we followed the original paper on supervised contrastive learning [27] to use 1000 epochs for CIFAR-LT, for both KCL and TSC. Here, we also compare with the results of training LDAM-DRW [4] for 1000 epochs on CIFAR-10-LT and CIFAT-100-LT ($\rho = 100$). As shown in Table 10, training for longer epochs does not change the results for LDAM-DRW [4].

In Sec. 4, we used 200 epochs for KCL because the original KCL paper uses 200 [23]. We used 400 epochs for TSC because half of TSC's epochs are for warm-up. Since TSC uses KCL for the warm-up, we started TSC after KCL has converged, i.e., after 200 epochs, and ran it for another 200 epochs. Here we also report the accuracy of TSC and KCL on ImageNet-LT for 200 and 400 epochs. As shown in Table 11, in both cases, TSC outperforms KCL.

Table 12. Performances of TSC , KCL and cross entropy loss on full CIFAR-10, CIFAR-100 and ImageNet.

Methods	CIFAR-10	CIFAR-100	ImageNet
CE	92.8	71.1	76.6
KCL†	93.0	70.5	77.0
TSC	92.9	70.8	77.1

Performance of TSC on balanced datasets: similar to KCL, TSC can also be applied to balanced datasets. In Table 12, we compare the performance of TSC vs. KCL and cross entropy loss on CIFAR-10, CIFAR-100 and ImageNet. As shown in the table, TSC achieves similar performance as KCL and cross-entropy loss on balanced datasets, and the improvements are far less than its improvements on imbalanced datasets. This further shows that TSC's improvements on imbalanced datasets comes from its ability to balance the feature space when the data distribution is imbalanced.

Table 13. TSC with different λ on CIFAR-10-LT with imbalance ratio 100.

λ	0.01	0.05	0.1	0.2	0.3	0.5	1	2	5	10
Acc (%)	78.4	79.2	79.6	**79.7**	79.4	79.5	78.8	78.5	78.2	78.0

Performance of TSC with different λ: In TSC, the targeted supervised contrastive loss is a weighted sum of two components, the first is a standard supervised contrastive loss as used in KCL [23], whereas the second is a contrastive loss between the target and the samples in the batch:

$$\mathcal{L}_{TSC} = -\frac{1}{N}\sum_{i=1}^{N}\Big(\frac{1}{k+1}\sum_{v_j^+\in \tilde{V}_{i,k}^+}\log\frac{e^{v_i^T\cdot v_j^+/\tau}}{\sum\limits_{v_j\in\tilde{V}_i\cup U} e^{v_i^T\cdot v_j/\tau}} + \lambda\log\frac{e^{v_i^T\cdot c_i^*/\tau}}{\sum\limits_{v_j\in\tilde{V}_i\cup U} e^{v_i^T\cdot v_j/\tau}}\Big)$$

In the experiments of main paper, λ is set to 0.2. In this section, we investigate how different λ affects the performance of TSC. Table 13 compares the performance of TSC with different λ on CIFAR-10-LT with imbalance ratio 100. As we can see from the results, TSC is robust to a wide range of values for λ. In particular, values between 0.1 and 0.5 all yield very good performance, with the best being 0.2. With smaller values, the performance drops slowly this is because a too small λ may not be enough to pull the samples to the targets. We also notice that with large λ, there is too much emphasis on pulling each class to the nearest target and less emphasis on keeping the classes that are semantically clause near each other. Therefore, we fix $\lambda = 0.2$ for all experiments.

Table 14. TSC with and without warm-up on ImageNet-LT.

Methods	Many	Medium	Few	All	**R**$\downarrow$
w/o warmup	62.1	48.9	29.3	51.3	7.65
w/ warmup	**63.5**	**49.7**	**30.4**	**52.4**	**7.14**

Warm-up Training. As mentioned in Sec. 4.1, we first warm up the network by not assigning targets and simply training the network with the KCL loss. As shown in Table 14, a network trained with a warm-up phase achieves much better accuracy and reasonability than without warm-up on ImageNet-LT. This is likely because in the early stage of training, the feature space is quite random. As a result, the class target assignment at such early stages is nearly random assignment and thus could prevent the feature space from learning good semantics.

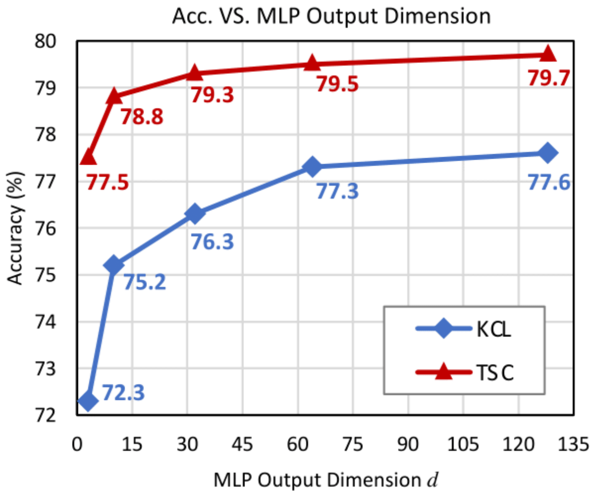

Figure 7. Comparison of KCL and TSC on different output feature dimensions d on CIFAR-10-LT with imbalance ratio 100.

Output Dimensions. One important hyper-parameter of contrastive learning is the output dimension of the MLP head. In all of our experiments, we set the default output dimension to be 128. In Fig. 7, we compare the performance of TSC and KCL with different MLP output dimensions on CIFAR-10-LT with imbalance ratio 100. As shown in the figure, the performance of TSC is quite robust to small output dimensions (77.5% at $d = 3$ and 79.6% at $d = 128$), while KCL experiences noticeable performance drop when the output dimension is small (only 72.3% at $d = 3$ and 77.6% at $d = 128$). This is possibly because when the output dimension is small, it is harder for KCL to achieve good uniformity, while TSC always achieves good uniformity because of the pre-computed targets.