Supplementary Material:

Regularizing Class-wise Predictions via Self-knowledge Distillation

A. Effects of hyper-parameters

To examine the effect of main hyper-parameters T and $\lambda_{\texttt{cls}}$, we additionally test the hyper-parameters across an array of $T \in \{0.1, 0.5, 1, 4, 10, 20\}$ and $\lambda_{\texttt{cls}} \in \{0.1, 0.5, 1, 2, 3, 4, 10, 20\}$ on PreAct ResNet-18 using the CIFAR-100 dataset. The results are presented in Table 1. Except for the hyper-parameters under consideration, we keep all settings the same as in Section 3.1. Overall, we found our method is fairly robust on T and $\lambda_{\texttt{cls}}$, except for some extreme cases, such as the small value of $T \leq 0.5$, and the large value of $\lambda_{\texttt{cls}} \geq 10$.

T \ $\lambda_{\texttt{cls}}$	0.1	0.5	1	2	3	4	10	20
0.1	25.16	24.03	23.91	24.38	24.05	24.21	24.39	27.61
0.5	24.14	24.05	24.15	23.49	23.78	23.23	23.90	25.96
1	24.15	23.32	22.80	22.26	22.87	23.18	24.35	25.58
4	22.87	22.03	**21.66**	22.45	22.68	22.81	32.25	35.45
10	22.68	22.36	21.98	22.04	21.95	31.76	31.80	37.50
20	22.96	22.39	22.03	22.37	22.00	22.39	30.23	24.05

Table 1. Top-1 error rates (%) of PreAct ResNet-18 on CIFAR-100 dataset over various hyper-parameters T and $\lambda_{\texttt{cls}}$. The best results are indicated in bold.

B. Qualitative analysis of CS-KD

To examine the effect of our method, we investigate prediction values in softmax scores, *i.e.*, $P(y|\mathbf{x})$, from PreAct ResNet-18 trained by the standard cross-entropy loss and our method for TinyImageNet dataset. We report commonly misclassified samples by both the cross-entropy and our method in Figure 1, and softmax scores of the samples show our method not only moderates the overconfident predictions, but also enhances the prediction values of classes correlated to the ground-truth class.

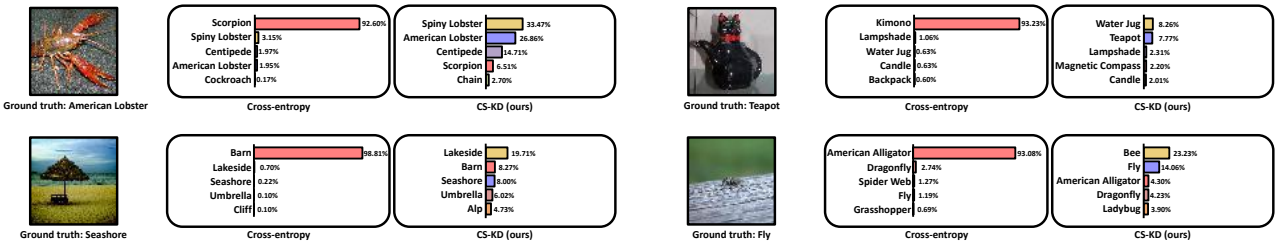
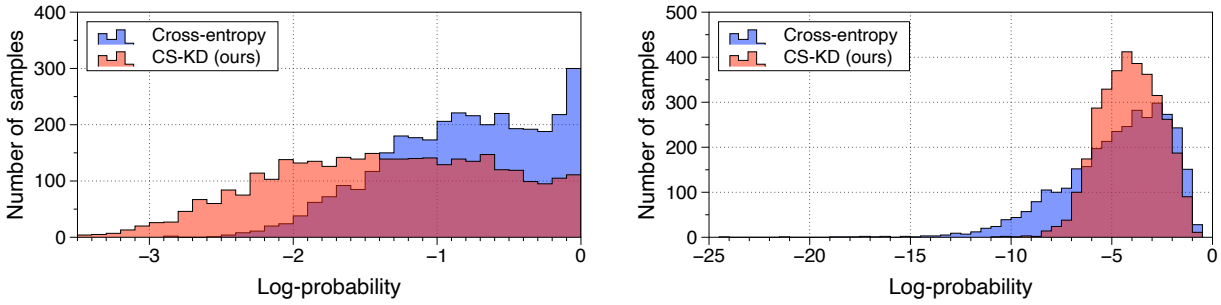


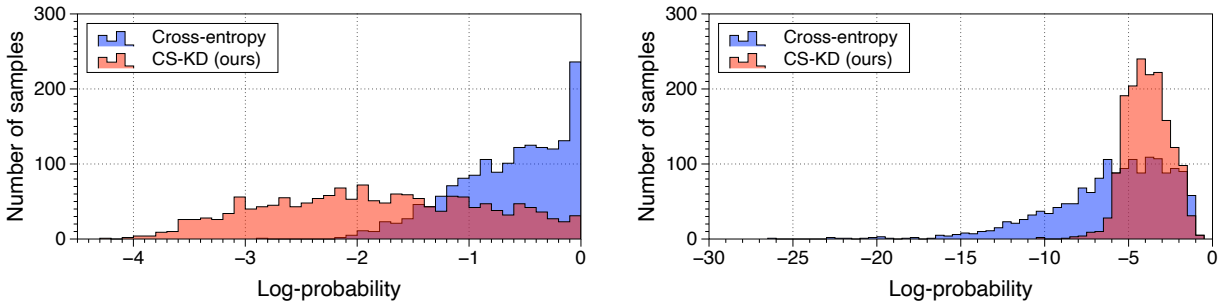
Figure 1. Predictive distributions on misclassified samples. We use PreAct ResNet-18 trained on TinyImageNet dataset. For misclassified samples, softmax scores of the ground-truth class are increased by training DNNs with class-wise regularization.

Moreover, we additionally compare our method with the cross-entropy method by plotting log-probabilities of the softmax scores on commonly misclassified samples for TinyImageNet, CUB-200-2011, Stanford Dogs, and MIT67 datasets. The corresponding results are reported in Figures 2, 3, 4, and 5. Log-probabilities of the softmax scores on the predicted class show how overconfident the predictions are, and our method produces less confident predictions on the misclassified samples compared to the cross-entropy method for overall datasets. On the other hand, log-probabilities of the softmax scores on the ground-truth class show relations between the predictions and the ground-truth class, and our method increases the ground-truth scores for overall datasets. These results imply that our method induces meaningful predictions that are more related to the ground-truth class than the cross-entropy method.



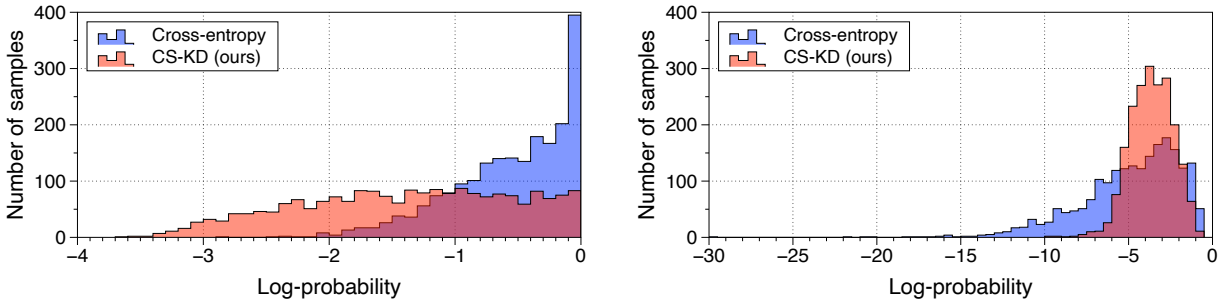
(a) Log-probabilities of predicted labels on misclassified samples (b) Log-probabilities of ground-truth labels on misclassified samples

Figure 2. Histogram of log-probabilities of (a) the predicted label, *i.e.*, top-1 softmax score, and (b) the ground-truth label on misclassified samples by networks trained by the cross-entropy (baseline) and CS-KD. The networks are trained on PreAct ResNet-18 for TinyImageNet.



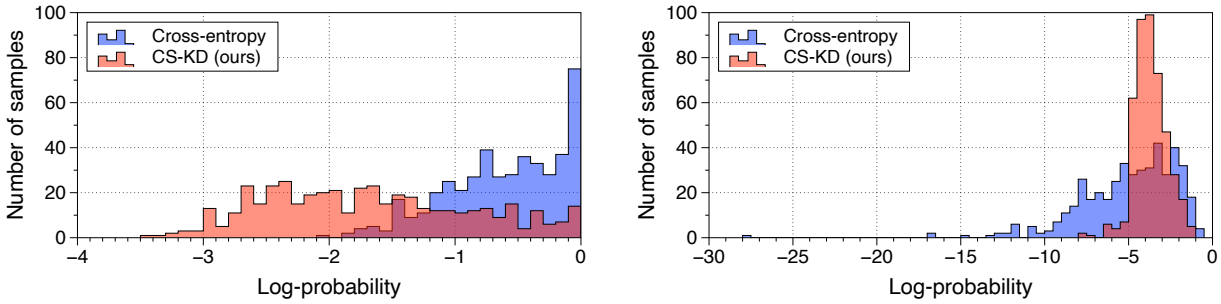
(a) Log-probabilities of predicted labels on misclassified samples (b) Log-probabilities of ground-truth labels on misclassified samples

Figure 3. Histogram of log-probabilities of (a) the predicted label, *i.e.*, top-1 softmax score, and (b) the ground-truth label on misclassified samples by networks trained by the cross-entropy (baseline) and CS-KD. The networks are trained on ResNet-18 for CUB-200-2011.



(a) Log-probabilities of predicted labels on misclassified samples (b) Log-probabilities of ground-truth labels on misclassified samples

Figure 4. Histogram of log-probabilities of (a) the predicted label, *i.e.*, top-1 softmax score, and (b) the ground-truth label on misclassified samples by networks trained by the cross-entropy (baseline) and CS-KD. The networks are trained on ResNet-18 for Stanford Dogs.



(a) Log-probabilities of predicted labels on misclassified samples (b) Log-probabilities of ground-truth labels on misclassified samples

Figure 5. Histogram of log-probabilities of (a) the predicted label, *i.e.*, top-1 softmax score, and (b) the ground-truth label on misclassified samples by networks trained by the cross-entropy (baseline) and CS-KD. The networks are trained on ResNet-18 for MIT67.