Table 8: Pre-trained Models. We list two variants of CLIP [43] that use ResNet-50 [15] and ViT-Base [10] as the image encoder respectively.

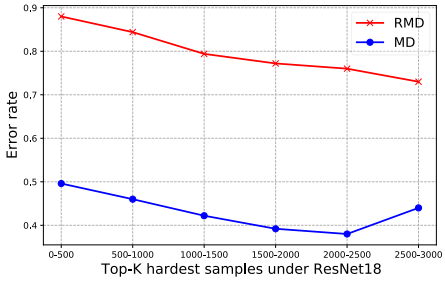
Model	Pre-trained dataset
CLIP-ViT-B CLIP-R50	400M image-caption data [43]
ViT-B	14M ImageNet21k (w. labels) [46]
MAE-ViT-B	1.3M ImageNet1k (w./o. labels) [7]

A The comparison of MD and RMD for measuring the sample difficulty

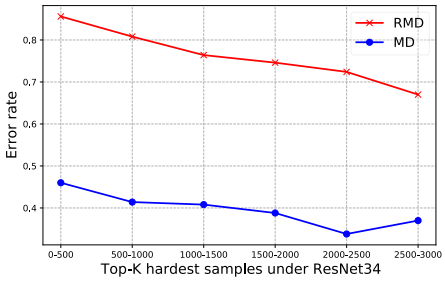
(a) Visualization of the Top-8 hardest samples (top row) and Top-8 easiest samples (bottom row) in ImageNet (class Tusker) which are ranked by means of the CLIP-VIT-B-based RMD score.

(b) Visualization of the Top-8 hardest samples (top row) and Top-8 easiest samples (bottom row) in ImageNet (class Tusker) which are ranked by means of the CLIP-VIT-B-based MD score.

Figure 5: Visualization of the Top-k hardest and easiest samples in ImageNet (class Tusker) which are ranked by RMD and MD scores. In contrast to the MD score, the easy and hard samples measured by RMD are more accurate than those by MD.



(a) Error rate achieved by ResNet18 (trained on ImageNet) on the validation subsets, which respectively contain 500 samples ranked from the ath to bth hardest.



(b) Error rate achieved by ResNet34 (trained on ImageNet) on the validation subsets, which respectively contain 500 samples ranked from the ath to bth hardest.

Figure 6: The performance comparison of RMD and MD for characterizing Top-K hardest samples.

In Fig. 5, we further compare Top-k hardest and easiest samples that are ranked by RMD (Fig. 5a) and MD (Fig. 5b) scores respectively. We can see that hard and easy samples characterized by RMD are more accurate than those characterized by MD, and there is a high-level agreement between human visual perception and RMD-based sample difficulty. Moreover, we quantitatively compare the

performance of RMD and MD for characterizing Top-K hardest samples in Fig 6. We can observe that the error rate of ResNet18 and ResNet34 on the hardest data split rated by RMD is close to 90%, which significantly suppresses the performance of MD. Therefore, the derived RMD in this paper is an improvement over the class-conditional MD for measuring the sample difficulty.

B Additional comparisons for different pre-trained models

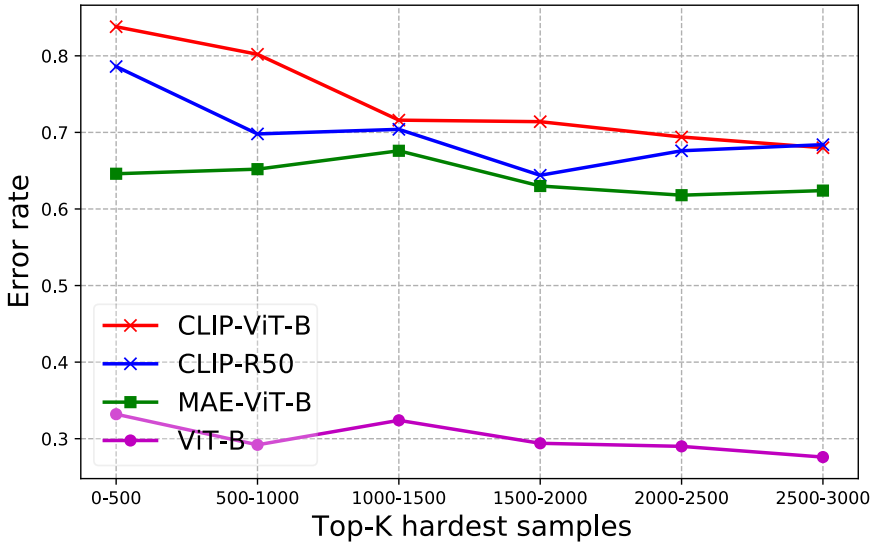
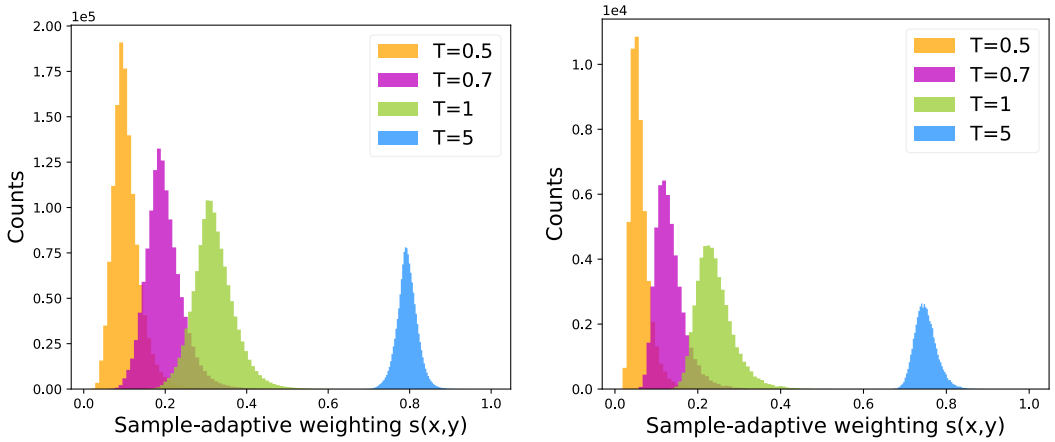


Figure 7: Error rate achieved by DenseNet121 (trained on ImageNet) on the validation subsets, which respectively contain 500 samples ranked from the ath to bth hardest. Four different pre-trained models are used for computing RMDs and ranking. They all show the same trend, i.e., the error rate reduces along with the sample difficulty. However, ViT-B supervisedly trained on ImageNet21k performed much worse than the others.

C More hard and easy samples ranked by RMD

Figure 8: Visualization of the Top-8 hardest samples (top row) and Top-8 easiest samples (bottom row) in ImageNet (class indigo bird) which are ranked by means of the CLIP-VIT-B-based RMD score.

Figure 9: Visualization of the Top-8 hardest samples (top row) and Top-8 easiest samples (bottom row) in ImageNet (class echidna) which are ranked by means of the CLIP-VIT-B-based RMD score.

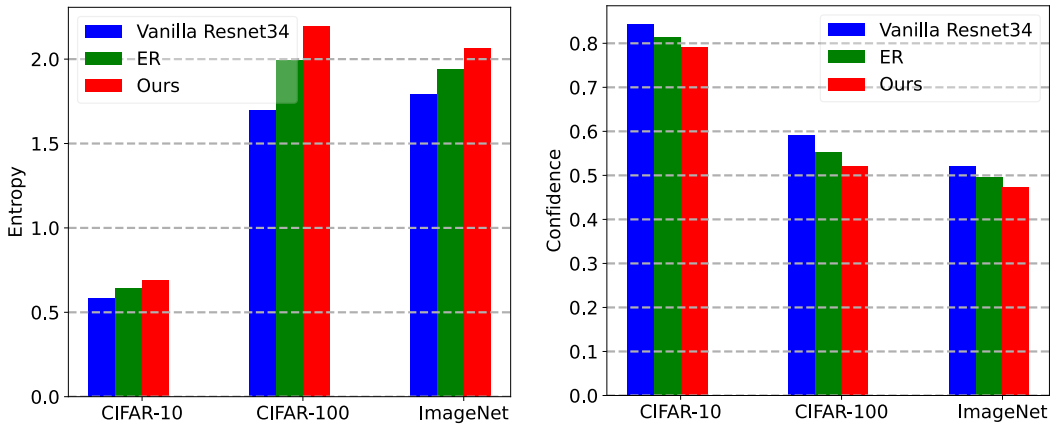


(a) ImageNet1k.

(b) CIFAR100.

Figure 10: Histograms of $s(x_i, y_i)$ at different T.

D More experimental results



(a) Predictive entropy of misclassified samples on CIFAR and ImageNet datasets.

(b) Predictive confidence of misclassified samples on CIFAR and ImageNet datasets.

Figure 11: Predictive entropy and confidence of misclassified samples for different methods on CIFAR and ImageNet datasets.

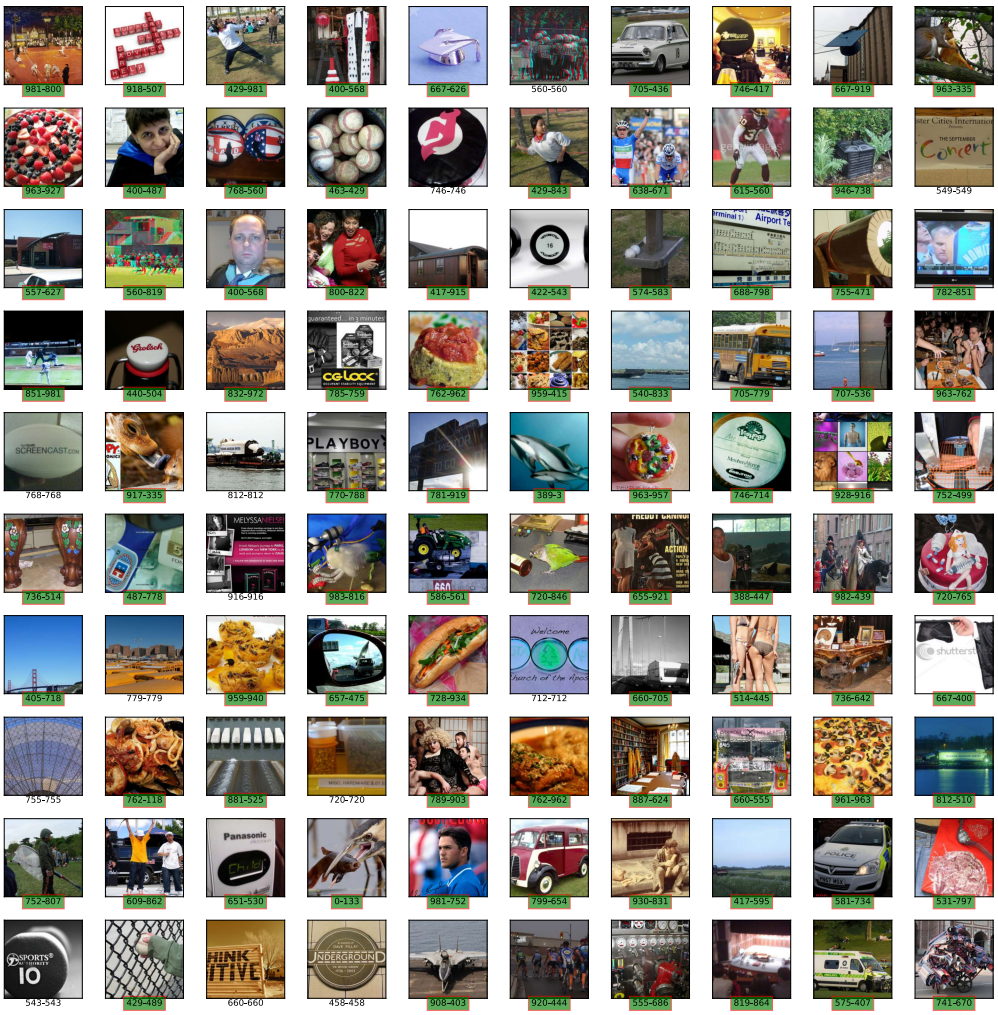


Figure 12: Test samples with predictions corresponding to the Top-100 relative Mahalanobis distance. The text box ("x-x") with green shade represents a misclassification. The first number indicates the label class, and the second number indicates the predictive class.

Table 9: The comparison of various architectures for predictive Top-1 accuracy (%) and ECE (%) on ImageNet1k.

Arch.		CE	ER	Proposed
ResNet18	ACC ↑	70.46	70.59	**70.82**
	ECE ↓	4.354	2.773	**1.554**
ResNet34	ACC ↑	73.56	73.68	**74.11**
	ECE ↓	5.301	3.720	**1.602**
ResNet50	ACC ↑	76.08	76.11	**76.59**
	ECE ↓	3.661	3.212	**1.671**
DenseNet121	ACC ↑	75.60	75.73	**75.99**
	ECE ↓	3.963	3.010	**1.613**
WRN50x2	ACC ↑	76.79	76.81	**77.23**
	ECE ↓	4.754	2.957	**1.855**

Table 10: The comparison of different model-based measures for predictive Top-1 accuracy (%) and ECE (%) on ImageNet1k. Compared to the three ResNets, "Δ" denotes the averaged gain achieved by CLIP-ViT-B in Table 5.

Measures		ResNet34	ResNet50	ResNet101	Δ
RMD	ACC $\uparrow$	73.73	73.78	73.88	$+0.31$
	ECE $\downarrow$	3.298	2.996	2.882	-1.44
Loss	ACC $\uparrow$	73.58	73.61	73.75	$+0.46$
	ECE $\downarrow$	3.624	2.997	2.783	-1.52