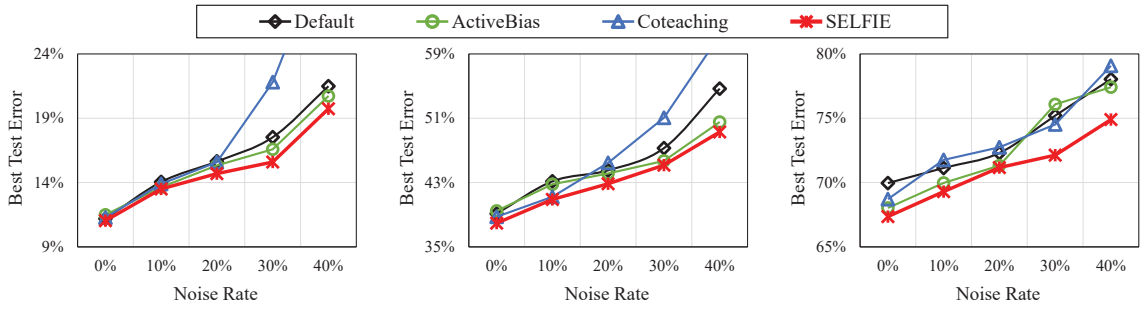
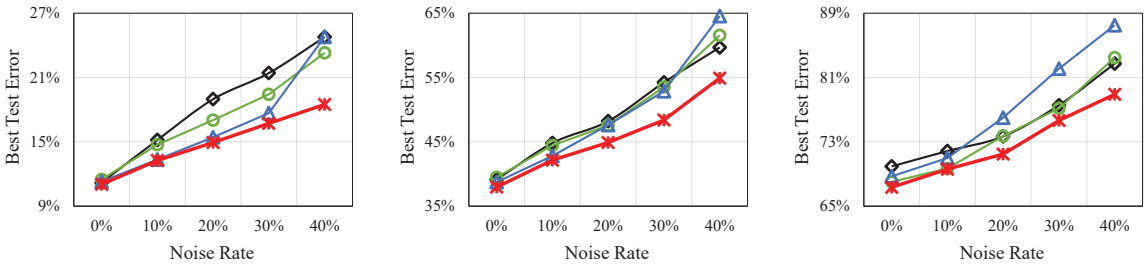


(a) CIFAR-10. (b) CIFAR-100. (c) Tiny-ImageNet.

Figure 11. The best test error of the four training methods using VGG-19 on three data sets with varying **pair noise** rates.



(a) CIFAR-10. (b) CIFAR-100. (c) Tiny-ImageNet.

Figure 12. The best test error of the four training methods using VGG-19 on three data sets with varying **symmetry noise** rates.

Table 5. The best test error (%) on **pair noise** 40% in Figure 11.

Method	CIFAR-10	CIFAR-100	Tiny-ImageNet
Default	21.5±1.37	54.7±0.13	78.0±0.26
ActiveBias	20.7±0.09	50.5±0.57	77.4±0.66
Coteaching	32.9±0.77	61.3±1.77	79.1±0.13
SELFIE	**19.7±0.18**	**49.3±0.10**	**74.9±0.11**

Table 6. The best test error (%) on **symmetry noise** 40% in Figure 12.

Method	CIFAR-10	CIFAR-100	Tiny-ImageNet
Default	24.8±0.19	59.7±0.17	82.7±0.60
ActiveBias	23.3±0.31	61.6±1.65	83.5±1.04
Coteaching	24.8±0.92	64.5±3.46	87.5±0.41
SELFIE	**18.5±0.13**	**54.9±0.38**	**78.9±0.33**

A. VGG-19 on Synthetic Noise

To validate the generality of the model, we trained VGG-19 (Simonyan & Zisserman, 2014) on the three synthetic data sets with the same configuration as in Section 4. Figures 11 and 12 show the test errors of the four training methods using VGG-19 with varying pair and symmetry noise rates. Again, *SELFIE* achieved the lowest test error at any noise rate of both noise types on all data sets. *SELFIE* outperformed the other methods by a larger margin at a higher noise rate. Especially when the noise rate was 40%, *SELFIE* reduced the *absolute* test error by 1.8*pp*–5.4*pp* compared with *Default*, 1.0*pp*–2.5*pp* compared with *ActiveBias*, and 4.2*pp*–13.2*pp* compared with *Coteaching* for the pair noise; by 3.8*pp*–6.3*pp* compared with *Default*, 4.6*pp*–6.7*pp* compared with *ActiveBias*, and 6.3*pp*–9.6*pp* compared with *Coteaching* for the symmetry noise. On the other hand, *ActiveBias* achieved the test error slightly lower than that of *Default*, and *Coteaching* worked well only on CIFAR-10 with symmetry noise, i.e., Figure 12(a). Tables 5 and 6 summarize the test errors of all methods at the noise rate of 40% for the pair and symmetry noises, respectively.

B. "ANIMAL-10N" for Realistic Noise

Because many researchers in robust optimization are facing a lack of real-world noisy data sets, we build a benchmark data set with realistic noises, which we call **ANIMAL-10N**, and publicly release it at https://dm.kaist.ac.kr/datasets/animal-10n in Figure 13.

Data Collection: To include human error in the image labeling process, we first defined five pairs of "confusing" animals: {(cat, lynx), (jaguar, cheetah), (wolf, coyote), (chimpanzee, orangutan), (hamster, guinea pig)}, where two animals in each pair look very similar. Then, we crawled 6,000 images for each of the ten animals on Google and Bing by using the animal name as a search keyword. Consequently, in total, 60,000 images were collected.

Data Labeling: For human labeling, we recruited 15 participants, which were composed of ten undergraduate and five graduate students, on the KAIST online community. They were educated for one hour about the characteristics of each animal before the labeling process, and each of them was asked to annotate 4,000 images with the animal names in

Table 7. Number of the images for each class in the training and test sets of ANIMAL-10N.

Class	Cat	Lynx	Wolf	Coyote	Cheetah	Jaguar	Chimpanzee	Orangutan	Hamster	Guinea pig	Total
Training Set	5466	4608	5091	4841	4981	4913	5322	4999	4970	4809	50000
Test Set	557	485	423	410	509	524	620	557	440	475	5000

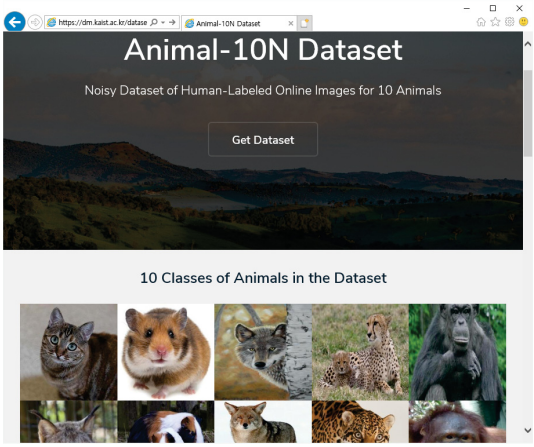


Figure 13. ANIMAL-10N homepage.

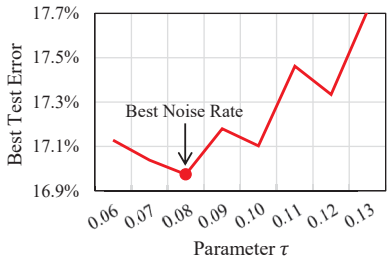


Figure 14. The best noise rate of the ANIMAL-10N data set obtained by grid search.

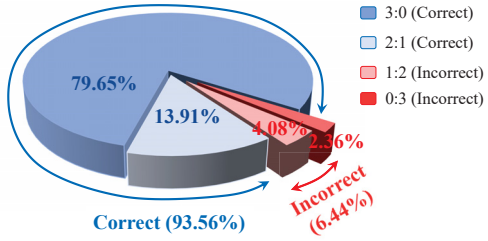


Figure 15. Proportion of correct and incorrect labels by human inspection.

a week, where an equal number (i.e., 400) of images were given from each animal. More specifically, we combined the images for a pair of animals into a single set and provided each participant with *five* sets; hence, a participant categorized 800 images as either of two animals five times. After the labeling process was complete, we paid about US$150 to each participant. Finally, excluding irrelevant images, the labels for $55,000$ images were generated by the participants. Please note that these labels may involve human mistakes because we intentionally mixed confusing animals.

Data Organization: We randomly selected $5,000$ images for the test set and used the remaining $50,000$ images for the training set. Because the test set should be free from noisy labels, only the images whose label matches the search keyword were considered for the test set. Besides, the images are almost evenly distributed to the ten classes (or animals) in both the training and test sets, as shown in Table 7.

Noise Rate Estimation by Accuracy: Because the ground-truth labels are unknown,[8] we estimated the noise rate τ by the cross-validation with grid search (Liu & Tao, 2016; Li et al., 2017). Following the same configuration as in Section 4, we trained DenseNet (L=25, k=12) using *SELFIE* on the $50,000$ training images and evaluated the performance on the $5,000$ testing images. As shown in Figure 14, the best noise rate was $\tau = 0.08$ from a grid $\tau \in [0.06, 0.13]$ when τ was incremented by 0.01. Therefore, we decided to set $\tau = 0.08$ for ANIMAL-10N.

Noise Rate Estimation by Human Inspection: We also estimated the noise rate τ by human inspection to verify the result based on the grid search. To this end, we randomly sampled 6,000 images and acquired two more labels for each of these images in the same way. Meanwhile, human experts different from the 15 participants carefully examined the 6,000 images to get the ground-truth labels. Comparing the human labels and the ground-truth labels in Figure 15, the former in the legend represents the number of the votes for the *true* label, and the latter represents the number of the votes for the other label. Because three votes were ready for each image, for conservative estimation, the final human label was decided by majority. Thus, the two cases of 3:0 and 2:1 were regarded as *correct* labeling, and the other two cases of 1:2 and 0:3 were regarded as *incorrect* labeling. Overall, the proportion of incorrect human labels was $4.08 + 2.36 = 6.44\%$ in the sample, and it is fairly close to $\tau = 0.08$ obtained by the grid search.

[8] One might think that the search keyword could be used as the true label, but we found that the search results contained non-negligible erroneous images.