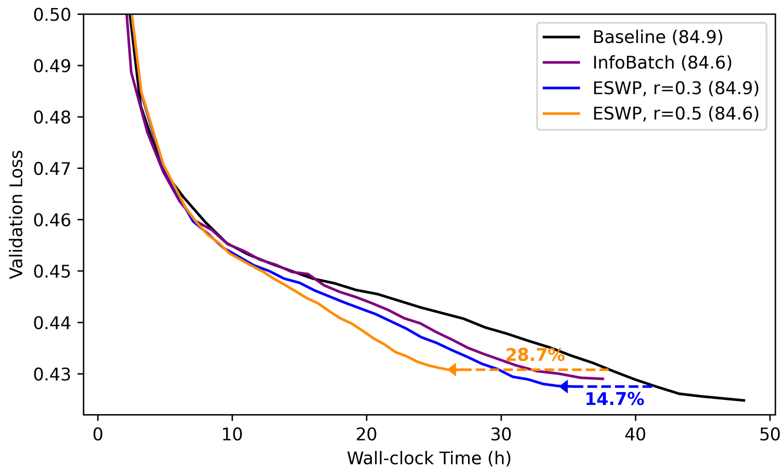

Figure 1: Re-construction losses of MAE pre-training ViT-Large on ImageNet-1K. The number in bracket in the legend is the validation accuracy (%) after fine-tuning. **ESWP achieves lossless acceleration over Baseline (no data selection), and consistently outperforms previous SOTA method InfoBatch.**

	Baseline	InfoBatch	ESWP (r=0.3)	ESWP (r=0.5)
Time(h)	48.1	37.6	35.1	**27.1**
Time saved(%)	-	21.8	27.0	**44.7**
Acc.(%)	84.9	84.6	**84.9**	84.6

Table 1: Comparisons of pre-training time and fine-tuning accuracy (Table 6 updated)

	Baseline	Random	ES	ESWP
Clean (0%)	81.1	$80.4_{\downarrow 0.7}$, 29%	$\mathbf{81.1}_{\uparrow 0.0}$, 11%	$80.6_{\downarrow 0.5}$, **31**%
Uniform (40%)	51.1	$52.9_{\uparrow 1.8}$, 20%	$\mathbf{60.1}_{\uparrow 9.0}$, 16%	$58.7_{\uparrow 7.6}$, **25**%

Table 2: Accuracy (%) and Time-Saved of ResNet-50 on CIFAR-100. Here Random renders Baseline with random data pruning, and **its performance is consistently worse than ESWP under the same amount of computation time saving.**

References

Truong Thao Nguyen, Balazs Gerofi, Edgar Josafat Martinez-Noriega, François Trahay, and Mohamed Wahib. KAKURENBO: Adaptively hiding samples in deep neural network training. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 37900–37922. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/7712b1075f5e0eae297702845714098f-Paper-Conference.pdf.

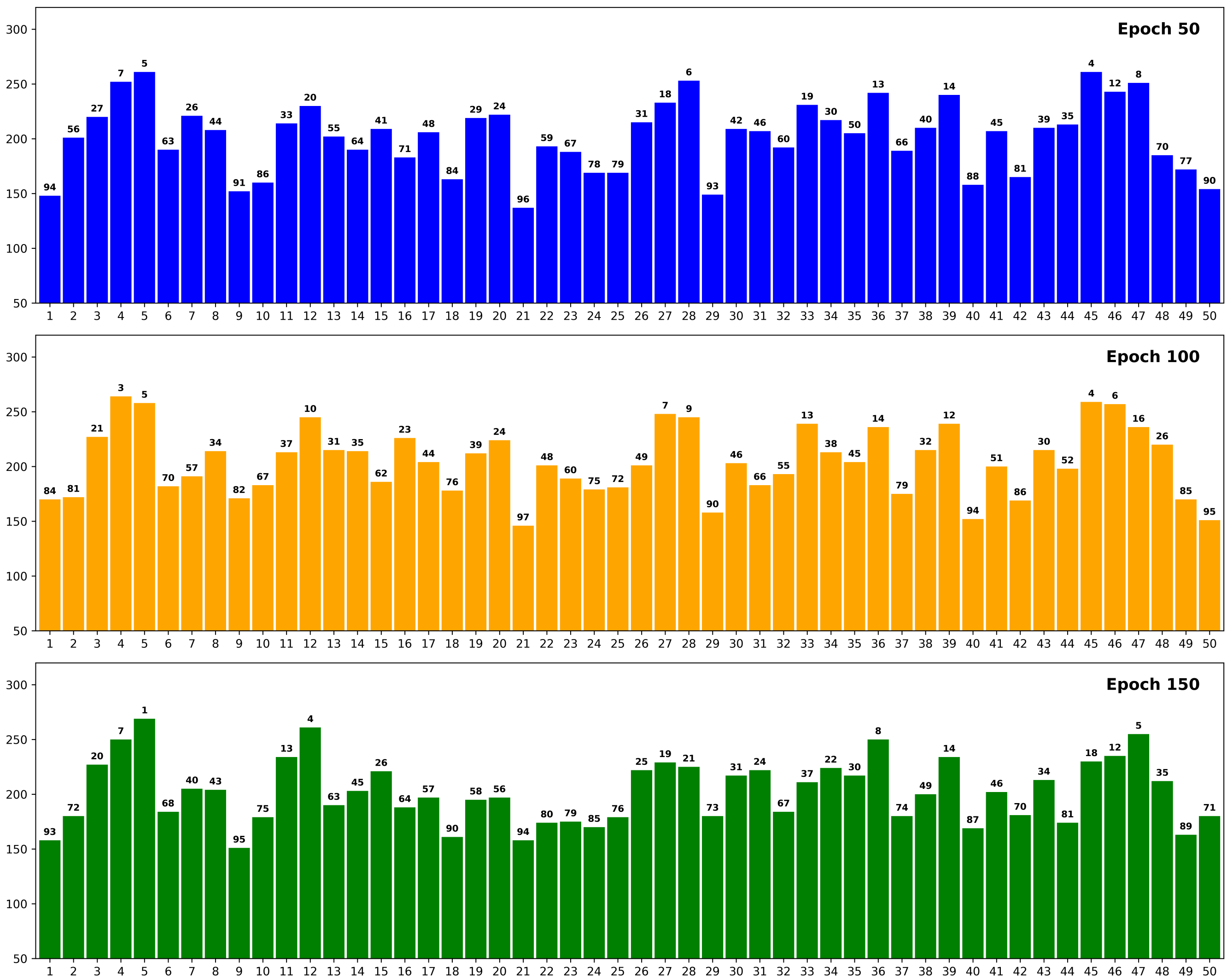

Figure 2: Visualization of the number of selected samples for BP of each class in ESWP (ResNet-50, Cifar-100), following Figure 6 in Thao Nguyen et al. [2023]. The figure shows the result of the first 50 classes. The number on top of each column shows the rank over 100 classes (a lower number indicates a higher number of selected samples). This indicates that ES(WP) can automatically select samples in different training stages.