What to Distill? Fast Knowledge Distillation with Adaptive Sampling

Supplementary Material

A. Hyperparameter Selection

KDAS uses two hyperparameters (i.e., the initial and final sampling ratios) for quantity-based subsampling. Once they are determined for a certain KD method such as vanilla KD and LogitSTD, they can be reused for different model architectures and datasets, as reported in the paper.

KDAS uses four hyperparameters (i.e., λ, γ, θ_{low}, θ_{high}) for quality-based calibration. We empirically tune these hyperparameters through a grid search on CIFAR-100. Table S1 presents an ablation study on the effects of the hyperparameters.

Table S1. Hyperparameter Exploration

λ	γ	θ_{low}	θ_{high}	VGG13 → VGG8	WRN40 → Res8×4
1000	0.5	20	80	**73.91**	**76.11**
1500	0.5	20	80	73.42	75.54
2000	0.5	20	80	73.38	75.41
1000	0.1	20	80	73.55	75.40
1000	0.3	20	80	73.89	75.63
1000	0.7	20	80	73.85	76.00
1000	0.5	10	80	73.88	76.04
1000	0.5	30	80	73.71	75.75
1000	0.5	40	80	73.44	75.55

The hyperparameters for quality-based calibration used in our experiments ($\lambda = 1000$, $\gamma = 0.5$, $\theta_{low} = 20$, $\theta_{high} = 80$) are found to work robustly across KD methods and model architectures.

B. Generalizability

B.1. Application to Vision Transformers

We apply KDAS to vision transformers in combination with a recent knowledge distillation method, LogitSTD. Table S2 shows the top-1 accuracy (%) of four vision transformer models on CIFAR-100 with ResNet56 as the teacher model.

Table S2. Application to Vision Transformers

Method	DeiT-Ti	T2T-ViT7	PiT-Ti	PVT-Ti
LogitSTD	78.55	78.43	78.76	78.43
LogitSTD + KDAS	77.43	77.98	**78.86**	77.63
Δ Accuracy	**-1.12%**	**-0.45%**	**+0.1%**	**-0.8%**
Δ Training Time	**-15.41%**	**-15.40%**	**-14.63%**	**-15.54%**

KDAS improves both accuracy and training efficiency for PiT-Ti only, implying that other transformer models may require more data to benefit from LogitSTD.

B.2. Application to Object Detection

We further apply KDAS to the object detection task with the PASCAL VOC dataset. We target the backbone network of an object detection model, Faster R-CNN, for distillation. Table S3 summarizes the accuracy and training time reductions for each teacher and student pair.

Table S3. Application to Object Detection (Metric: mAP)

T → S	KD	KD + KDAS	DKD	DKD + KDAS
R101 → R18	39.23	**39.97** (**-9.09%**)	38.04	**38.32** (**-9.06%**)
R50 → MV2	36.14	36.13 (**-9.10%**)	35.15	**35.91** (**-9.06%**)

The results demonstrate a broader applicability of KDAS beyond the classification task.

C. Comparison of Different Sampling Metrics

To justify the choice of KL divergence for sampling, we compare alternative metrics (i.e., Jensen–Shannon (JS) divergence and cross-entropy) for quantity-based subsampling in KDAS. Table S4 presents the classification accuracy across various architectures on CIFAR-100, under a fixed sampling ratio of 50%.

Table S4. Comparison of Different Sampling Metrics

Teacher **Student**	WRN-16-2 ResNet8x4	VGG13 VGG8	ResNet110 ResNet20
KL Divergence	**76.11**	**73.91**	**70.72**
JS Divergence	76.04	73.86	70.57
Cross-Entropy	75.98	73.40	70.15

D. Comparison with Data-Centric KD

We compare KDAS with recent data-centric distillation methods (KCD and UNIXKD) on CIFAR-100 for three architectures. Table S5 reports the accuracy and training time reductions obtained by each method.

Table S5. Comparison with Data-Centric KD Methods

Teacher **Student**	VGG13 VGG8	ResNet56 ResNet20	ResNet50 MobileNetV2s
KDAS	**73.91** (**-30.22%**)	**71.66** (**-27.1%**)	**68.25** (**-27.31%**)
KCD [14]	73.44 (**-18.4%**)	70.75 (**-18.4%**)	67.94 (**-18.4%**)
UNIXKD [5]	73.18 (**-23.65%**)	70.06 (**-23.99%**)	67.11 (**-24.69%**)

The results show that KDAS outperforms KCD and UNIXKD, achieving higher performance and greater reductions in training time, demonstrating the effectiveness of KDAS in data-centric KD scenarios.