Appendix: Sample-level Adaptive Knowledge Distillation for Action Recognition

Ping Li, Chenhao Ping
Hangzhou Dianzi University
Hangzhou, China
lpcs@hdu.edu.cn,pch@hdu.edu.cn

Wenxiao Wang, Mingli Song
Zhejiang University
Hangzhou, China
wenxiaowang@zju.edu.cn,brooksong@zju.edu.cn

ACM Reference Format:
Ping Li, Chenhao Ping and Wenxiao Wang, Mingli Song. 2025. Appendix: Sample-level Adaptive Knowledge Distillation for Action Recognition. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3746027.3755514

Notes: This supplementary primarily contains the following parts, including datasets, evaluation metric, compared methods, model size & computational cost, quantitative results on ImageNet and that with pre-training on ImageNet rather than Kinetics-400, the ablations on the interruption threshold η, the distillation strength trade-off parameter λ, the interruption rate function $\beta(\cdot)$, the sample selection interval, the sample distillation strength α, the average sample distillation strength α, the supervisor Teacher or Ground-Truth (GT) label, and the diversity trade-off parameter γ, as well as the visualization on the trend of selected sample number change per class across different epochs.

All experiments were performed on a server equipped with four 11G GeForce 2080Ti graphics cards. The codes are compiled with PyTorch 1.7, Python 3.8, and CUDA 10.1. The code implementation is available at https://github.com/mlvccn/SAKD_ActionRec.

1 DATASETS AND EVALUATION METRIC

We conduct experiments on three video benchmarks including UCF101 [12], Kinetics-400 [7], and Something-Something v2, as well as two image benchmarks including CIFAR-100 [8] and ImageNet [2]. Note that the results on ImageNet are only shown in this Appendix. Dataset statistics are shown in Table 1.

Table 1: Statistics of data. "K"/"N" is class/sample number.

Dataset	Training		Validation		Test	
	K	*N*	*K*	*N*	*K*	*N*
UCF101[12]	101	9,537	-	-	101	3,783
Kinetics-400[7]	400	234,619	400	19,761	-	-
Sth-Sth v2[4]	174	168,913	174	24,777	174	27,157
CIFAR-100[8]	100	50,000	-	-	100	10,000
ImageNet[2]	1,000	1,200,000	1,000	50,000	1,000	100,000

UCF101[1] [12] consists of daily-life action videos collected from YouTube, covering 101 different categories with a total of 13,320 video clips, whose total duration is approximately 27 hours. The videos are organized into 25 groups by category, with each group containing 4 to 7 videos. The video resolution is 320×240, and the frame rate is 25 fps. We use the official splits with all images uniformly cropped to a size of 224×224.

Kinetics-400[2] [1] was originally released by DeepMind and contains video clips collected from YouTube, covering 400 different action categories. Each category has at least 400 videos, with each video clip lasting approximately 10 seconds. It includes 234,619 samples for the training set and 19,761 videos for the validation set, with all input images uniformly cropped to a size of 256 × 256.

Something-Something v2[3] [4] is a collection of 220,847 labeled video clips of humans performing pre-defined, basic actions with everyday objects. Among these videos, there are 168,913 ones for training, 24,777 for validation, and 27,157 for test. It involves 174 action categories with the video duration ranging from 2 to 6 seconds. The average number of videos per class is 620, and the average duration of videos is 4.03 seconds. Most video frames have a resolution of 240× 320 pixels, and they are uniformly cropped to the size of 224×224.

CIFAR-100[4] [8] contains 100 categories, each of which has 600 RGB images of size 32×32. The training set has 50,000 images and the test set has 10,000 images. All images are uniformly rescaled to 32×32.

ImageNet[5] [2] is a subset of ILSVRC-2012, which contains approximately 1.2 million training images, 50,000 validation images, and 100,000 test images, covering 1,000 categories with about 1,200 images per category. The image resolutions vary, and they are cropped to the size of 224×224.

Following previous works [9, 16, 20], we adopt the commonly used Top-1 accuracy and Top-5 accuracy as the evaluation metrics. Top-1 or Top-5 accuracy evaluates those samples whose ground-truth class takes up the top or or one of the five leading positions of the candidate class set. Also we report the elapsed time of each epoch to show the training efficiency.

2 IMPLEMENTATION SETUP SUPPLEMENT

The batch size of our method using frozen model is different from that of student training. We have shown the batch size and the maximum epoch for two video datasets in Table 2. Given a test video or image, we do the normalization before feeding them into

MM '25, October 27–31, 2025, Dublin, Ireland.
© 2025 Association for Computing Machinery.
ACM ISBN 979-8-4007-2035-2/2025/10...$15.00
https://doi.org/10.1145/3746027.3755514

[1] https://www.crcv.ucf.edu/data/UCF101.php
[2] https://deepmind.com/research/open-source/kinetics
[3] https://www.qualcomm.com/developer/software/something-something-v-2-dataset
[4] http://www.cs.toronto.edu/ kriz/cifar.html
[5] https://image-net.org/index.php

Table 2: Batch size and training epoch.

Model	Batch Size						N_{epoch}		
	Our method			Student training					
	UCF101	Sth-Sth	Kinetics	UCF101	Sth-Sth	Kinetics	UCF101	Sth-sth	Kinetics
SlowFast [3]	96	128	128	32	48	48	50	200	200
TPN [17]	24	32	32	4	8	8	50	200	200
Video ST [10]	24	48	48	4	8	8	100	200	200

the student model to output the estimated action class or image label. The settings for ImageNet [2] are the same as that for CIFAR-100.

Due to the limited resource, the number of sampled frames for student is $T = 16$ on Kinetics-400 [7], and that for teacher is set to the same. Note that the performance will be greatly upgraded when more frames are sampled in each video clip. For example, when $T = 64$, Top-1 accuracy is 76.90% with SlowFast and 80.60% with VideoST. Although the favourable improvements are attained, the training time is also largely lengthened, e.g., it needs nearly 7 hours to train SlowFast for each epoch on a machine with four 2080Ti graphics cards.

3 COMPARED METHODS

We compare our SAKD method with two groups of State-Of-The-Art (SOTA) KD methods: 1) *logit-based* ones include KD (vanilla Knowledge Distillation) [6] [6], DKD (Decoupled KD) [19] [7], and CTKD (Curriculum Temperature KD) [9] [8]; 2) *feature-based* ones include CrossKD (Cross-head KD) [14] [9], DualKD [15], and GKD (Generative model based KD) [13] [10]. We use the source codes publicly available from the original papers, and we try the best to implement DualKD [15] by ourself since its code is unavailable. Our code is available in the attached file.

To verify the generalization ability of our SAKD method on both video and image samples, we examine the compared KD methods on three typical action recognition models, including SlowFast [3] [11], TPN (Temporal Pyramid Network) [17] [12], and VideoST (Video SwinTransformer) [10] [13], as well as two typical image classification models (settings follows [11]), including ResNet (Residual Neural Network) [5] [14] and WRN (Wide ResNet) [18] [15]. For SlowFast, we sample 16 or 8 or 4 frames in a clip when the step is set to 8 or 16, termed SF16x8 or SF8x8 or SF4x16. For TPN teacher or student, we sample 32 or 8 frames along the temporal dimension and scale up frames along the spatial dimension by a factor of 2 or 8, termed TPN-f32s2 or TPN-f8s8. For VideoST, SwinS is teacher whose total Transformer layer at Stage 3 is 18, while SwinT is student whose that layer number is 6. For SlowFast and TPN, teacher and student adopt ResNet101 and ResNet50 as the backbone respectively.

Table 3: Computational cost and speed of basic models.

Basic Model	Backbone	Tea.	Stu.	Params(M)↓	GFLOPs↓	FPS↑
TPN [17]	TPN-f32s2	✓		99.71	375.09	298
	TPN-f8s8		✓	71.80	202.05	516
SlowFast [3]	SF16×8	✓		62.14	97.19	953
	SF8×8	✓		34.57	50.86	1482
	SF4×16		✓	33.79	28.01	2125
VideoST [10]	SwinS	✓		49.10	138.14	361
	SwinT		✓	27.81	70.91	597
WideRN [18]	WRN40-2	✓		2.26	0.33	267
	WRN40-1		✓	0.57	0.08	202
	WRN16-2		✓	0.70	0.10	515
ResNet [5]	RN110	✓		1.74	0.26	104
	RN56	✓		0.86	0.13	204
	RN32×4	✓		7.43	1.09	324
	RN32		✓	0.47	0.07	249
	RN20		✓	0.28	0.04	384
	RN8×4		✓	1.23	0.18	722

Table 4: Results on the ImageNet [2] validation set. "r" is sample selection ratio.

Method	Type	r	ResNet34→ResNet18					
			Top1↑	Δ	Top5↑	Δ	hrs↓	Δ
-	Teac.	100%	73.31	-	91.58	-	1.6	-
-	Stud.	100%	69.75	-	89.43	-	1.3	-
KD[6] arxiv'15	Vani.	100%	70.34	-	90.31	-	1.4	-
	Ours	50%	**71.07**	+0.73	**90.92**	+0.61	1.2	-0.2h
	Ours*	50%	70.52	+0.18	90.83	+0.52	**0.9**	-0.5h
DKD [19] CVPR'22	Vani.	100%	71.51	-	90.25	-	1.5	-
	Ours	50%	**72.36**	+0.85	**90.87**	+0.62	1.3	-0.2h
	Ours*	50%	71.82	+0.31	90.34	+0.09	**0.8**	-0.7h
CTKD [9] AAAI'23	Vani.	100%	71.32	-	90.27	-	1.5	-
	Ours	50%	**72.87**	+1.55	**91.02**	+0.75	1.2	-0.3h
	Ours*	50%	71.72	+0.40	90.45	+0.18	**0.9**	-0.6h
GKD[13] AAAI'24	Vani.	100%	69.82	-	89.35	-	1.6	-
	Ours	50%	**71.15**	+1.33	**89.92**	+0.57	1.3	-0.3h
	Ours*	50%	70.34	+0.52	89.57	+0.22	**1.0**	-0.6h
CrossKD [14] AAAI'24	Vani.	100%	70.12	-	89.43	-	1.4	-
	Ours	50%	**71.29**	+1.17	**90.05**	+0.62	1.2	-0.2h
	Ours*	50%	70.85	+0.73	89.78	+0.35	**0.9**	-0.5h
DualKD [15] TIP'24	Vani.	100%	68.92	-	88.97	-	1.7	-
	Ours	50%	**70.27**	+1.35	**89.65**	+0.68	1.2	-0.5h
	Ours*	50%	69.18	+0.26	89.12	+0.15	**1.0**	-0.7h

4 MODEL SIZE AND COMPUTATIONAL COST

To help readers understand the performance gains brought by knowledge distillation methods, we show the model size and the computational cost of basic models for action recognition or image classification in Table 3. Here, the model size is evaluated in terms of Million Parameters (MParams), the model size is evaluated in terms of GFLOPs, and the inference speed is evaluated in terms of FPS (Frame Per Second). Among the video models, SlowFast [3] is the smallest one at the fastest inference speed.

[6] https://github.com/labmlai/annotated_deep_learning_paper_implementations
[7] https://github.com/megvii-research/mdistiller
[8] https://github.com/zhengli97/ctkd
[9] https://github.com/jbwang1997/crosskd
[10] https://github.com/aaai-24/generative-based-kd
[11] https://github.com/facebookresearch/SlowFast
[12] https://github.com/decisionforce/TPN
[13] https://github.com/SwinTransformer/Video-Swin-Transformer
[14] https://github.com/fastai/fastai
[15] https://github.com/szagoruyko/wide-residual-networks

Table 5: Performance on UCF101 [12] in terms of Top-1/5 Accuracy (%) / training time (min). "r" is sample selection ratio and without pretraining on Kinetics-400.

Method	Type	r	SlowFast[3]			VideoST[10]			TPN[17]		
			Top1↑	Top5↑	min↓	Top1↑	Top5↑	min↓	Top1↑	Top5↑	min↓
-	Teac.	100%	90.23	97.22	10.2	86.99	97.12	22.2	88.95	96.23	32.8
	Stud.	100%	83.13	95.72	6.2	84.40	96.17	18.1	80.03	93.83	20.7
KD[6] arXiv'15	Vani.	100%	84.20	95.22	6.4	85.69	96.63	20.0	81.30	94.24	23.4
	Ours	10%	**84.87**	**96.60**	4.6	**86.49**	**97.02**	11.1	**84.19**	**95.42**	11.3
	Ours*	10%	84.39	96.41	**1.4**	86.10	96.92	**3.9**	83.27	95.28	**4.9**
DKD[19] CVPR'22	Vani.	100%	85.06	96.09	6.8	84.92	96.83	19.9	82.17	93.98	23.5
	Ours	10%	**85.37**	**96.86**	4.6	**85.51**	**97.47**	12.2	**85.01**	**96.08**	13.0
	Ours*	10%	84.42	96.57	**1.5**	85.32	97.28	**4.0**	84.27	95.74	**5.0**
CTKD[9] AAAI'23	Vani.	100%	84.71	95.82	6.9	85.40	96.32	20.1	82.93	94.28	24.2
	Ours	10%	**85.31**	**96.13**	4.8	**86.24**	**97.09**	12.1	**84.28**	**96.23**	11.5
	Ours*	10%	85.02	95.91	**1.4**	85.98	96.73	**4.0**	83.29	95.85	**4.9**
GKD[13] AAAI'24	Vani.	100%	84.36	95.23	7.0	84.99	96.04	20.8	83.24	93.57	24.9
	Ours	10%	**84.92**	**96.84**	5.3	**85.93**	**96.98**	13.8	**84.28**	**95.38**	13.9
	Ours*	10%	84.62	95.63	**1.8**	85.28	96.50	**4.6**	83.49	95.02	**4.8**
CrossKD[14] CVPR'24	Vani.	100%	84.39	95.37	6.5	85.86	96.32	19.8	85.35	96.61	23.2
	Ours	10%	**84.89**	**95.93**	4.9	**86.93**	**97.09**	11.6	**86.38**	**97.41**	12.0
	Ours*	10%	84.18	95.31	**1.5**	86.36	96.92	**3.8**	85.72	96.98	**4.9**
DualKD[15] TIP'24	Vani.	100%	84.50	95.62	7.0	85.82	96.32	20.5	83.41	93.24	26.2
	Ours	10%	**85.32**	**96.32**	5.0	**86.76**	**97.05**	13.0	**84.04**	**95.78**	11.4
	Ours*	10%	85.19	96.27	**1.6**	86.01	96.83	**4.3**	83.75	95.34	**5.0**

Table 6: Ablation studies on the interruption threshold η with CrossKD [14].

η	UCF101 [12]				CIFAR-100 [8]			
	Top1 ↑		Top5 ↑		Top1 ↑		Top5 ↑	
	SlowFast	VideoST	SlowFast	VideoST	ResNet	WRN	ResNet	WRN
0.0	88.13	85.09	98.15	96.73	67.76	72.09	90.15	92.23
0.1	88.56	85.72	98.47	97.12	68.26	72.22	90.62	92.12
0.3	89.08	86.02	98.62	**97.52**	**69.28**	**73.92**	92.97	93.24
0.5	**90.04**	**86.94**	**99.23**	97.16	69.21	73.87	92.93	**93.87**
0.7	89.74	86.59	98.38	97.01	68.99	72.76	**93.28**	93.65
0.9	89.32	86.04	98.12	96.72	68.19	72.08	91.57	92.41
1.0	89.35	85.93	98.14	96.69	67.76	71.48	90.86	92.12

Table 7: Ablation studies on the distillation strength λ with CrossKD [14].

λ	UCF101 [12]				CIFAR-100 [8]			
	Top1 ↑		Top5 ↑		Top1 ↑		Top5 ↑	
	SlowFast	VideoST	SlowFast	VideoST	ResNet	WRN	ResNet	WRN
0.0	89.54	86.03	99.03	96.83	68.93	74.09	92.43	93.42
0.1	**90.04**	86.94	**99.23**	**97.16**	**69.21**	**73.87**	**92.93**	**93.87**
0.3	89.30	**87.03**	98.80	96.96	68.98	73.84	91.87	93.52
0.5	89.52	86.62	98.23	96.72	67.73	71.55	91.83	92.88
0.7	89.46	86.52	98.83	96.48	67.82	71.25	91.82	92.76
0.9	89.12	86.32	98.87	96.34	67.19	71.08	91.56	92.37
1.0	88.82	86.24	98.27	96.21	67.59	71.23	91.63	92.14

5 QUANTITATIVE RESULTS SUPPLEMENT

Results on ImageNet. Besides the reported results on CIFAR-100 (See Table 5 in the paper), we report the classification results on a large-scale image database, i.e., ImageNet, in Table 4. From the table, we observe that our SAKD method outperforms vanilla KD consistently when being applied to several SOTA KD methods at a much lower training cost. In particular, the performance gains are larger when selecting samples every epoch compared to that with every five epochs, because the former helps the student to learn more knowledge from more diverse samples.

Pre-training on ImageNet. As a common practice, the action recognition models adopt the model weights by pre-training the model on Kinetics-400. Here, we examine the KD performance when the action recognition model adopts the weights by pre-training the model on ImageNet rather than Kinetics-400. The results are shown in Table 5, which shows that our KD method still enjoys promising performances across several SOTA action recognition models at much lower training cost. Meanwhile, we observe that the performance decays are not large between ImageNet and Kinetics-400 (See Table 2 in the paper) on vanilla teacher and student models, e.g., 2% to 3% in terms of Top-1 Accuracy. This validates the importance of temporal relations in video, and the powerful representation ability of the model pre-trained on ImageNet.

6 ABLATION STUDIES SUPPLEMENT

Some more ablation studies on the interruption threshold η, the diversity trade-off parameter γ, the interruption rate function $\beta(\cdot)$, the sample selection interval, the sample distillation strength α, the average sample distillation strength α, the supervisor Teacher or Ground-Truth (GT) label, and the distillation difficulty-diversity trade-off parameter γ, were conducted on UCF101 and CIFAR-100. Here, we use ResNet56/ResNet20 and WRN40-2/WRN16-2 as teacher/student for CIFAR-100. We supplement the ablations on η and λ using CrossKD [14] in Table 6 and Table 7, respectively, while that results using KD [6] are in the full paper. The other ablations provide the results of both KD (logit-based) and CrossKD (feature-based) by selecting samples every five epochs, and hyper-parameters keep still as during training unless specified.

Interruption threshold η. We vary η from 0 to 1 with seven grids, and show the results of CrossKD in Table 6. From the table, we observe the similar tendency with that of KD in the paper, i.e., the performance rises up at early stage when the threshold starts from 0, and achieves the best around 0.5 in most situations across UCF101 and CIFAR-100. Then, the performance deteriorates at later stage when the threshold continues to increase. This suggests that when the interruption threshold should be neither too large nor too small, and it determines whether conducting the random dropout ($< \eta$) or the random shuffle on video frames ($\geq \eta$).

Distillation strength parameter λ. We vary λ from 0 to 1 with seven grids, and show the results of CrossKD in Table 7. From the table, we observe the similar tendency with that of KD in the paper, i.e., the performance achieves the best when λ takes 0.1. This means the previous distillation strength has 10% of the contribution, while the current one has 90% contribution to the model. That is to say, we should also consider the history distillation strength when the current one dominates the adaptive distillation process.

Interruption rate function $\beta(\cdot)$. We adopt three kinds of interruption rate function, including polynomial, linear, and cosine, whose results are shown in Table 8 (KD) and Table 9 (CrossKD). From the table, we observe the polynomial function achieves the best among the three, and the influences of the function are larger on CIFAR-100 than on UCF101. This means the interruption rate function form seems more robust on video than on image. Moreover,

Table 8: Ablation studies on the interruption rate function $\beta(\cdot)$ with KD [6].

$\beta(n)$	UCF101 [12]				CIFAR-100 [8]			
	Top1↑		Top5↑		Top1↑		Top5↑	
	SlowFast	VideoST	SlowFast	VideoST	ResNet	WRN	ResNet	WRN
Poly	**89.88**	**87.70**	**98.80**	**98.72**	**70.55**	**74.23**	**92.85**	**93.79**
Linear	89.12	86.28	98.70	97.38	68.73	72.97	91.66	92.85
Cosine	89.31	86.59	98.56	97.26	68.59	72.59	91.65	92.79

Table 9: Ablation studies on the interruption rate function $\beta(\cdot)$ with CrossKD [14].

$\beta(n)$	UCF101 [12]				CIFAR-100 [8]			
	Top1 ↑		Top5 ↑		Top1 ↑		Top5 ↑	
	SlowFast	VideoST	SlowFast	VideoST	ResNet	WRN	ResNet	WRN
Poly	**90.04**	**86.94**	**99.23**	**97.16**	**69.21**	**73.87**	**92.93**	**93.87**
Linear	89.40	86.01	98.64	96.92	68.02	71.24	91.72	92.31
Cosine	89.56	85.98	98.43	96.72	67.54	71.98	91.32	92.65

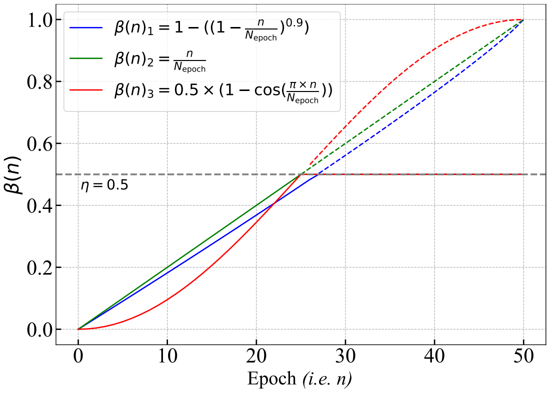
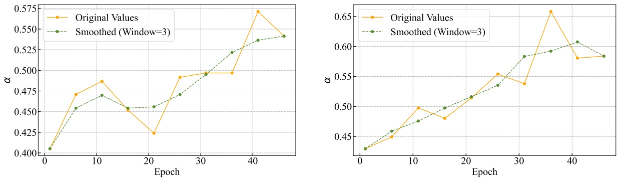


Figure 1: Different $\beta(n)$ vs epoch.

we depict the curve of $\beta(\cdot)$ under increasing epochs on UCF101 with KD in Fig. 1. As show in this figure, the outputs of all three kinds of functions rise up linearly or almost linearly with an increasing epoch n until it reaches the threshold $\eta = 0.5$. From the three curves, the random dropout operations are conducted with more epochs by adopting the polynomial function compared to the other two, and the dropout ratio is lower than that using linear function and higher that using cosine function during the first 20 epochs. This indicates the number of the dropout frames in a batch should be reasonable and moderate.

Sample selection interval. We vary the sample selection interval from 1 to 10 with five grids, and show the results in Table 10 (KD) and Table 11 (CrossKD). From the table, it can be seen that the performance in terms of Top-1 accuracy is consistently degenerated when the sample selection interval increases, while the elapsed training time reduces, i.e., the performance is better when we select samples at a smaller interval of epochs. But this is not always true for the case in terms of Top-5 accuracy, since there exist some fluctuations when selecting samples every one, three, or five epochs. Hence, we report the results of the interval takes both 1 and 5, where the latter is a good choice to strike the balance between the promising performance and the fast training speed.



(a) KD [6] **(b) CrossKD [14]**

Figure 2: Distillation strength α vs epoch.

Supervisor GT/Teacher. We explore the performance when only using teacher or Ground-Truth (GT) label to supervised the student learning, and show the results in Table 12 (KD) and Table 13 (CrossKD). From the table, we observe that both teacher and GT labels act as supervisor lead to the similar performance, while the teacher performs slightly better than GT labels. This demonstrates that both teacher and GT play important roles in guiding the student training, and the teacher has rich knowledge from the pre-trained model which is beneficial for boosting the performance of action recognition or image classification.

Diversity parameter γ. We vary the sample diversity parameter γ from 0 to 1 with seven grids, and show the results in Table 14 (KD) and Table 15 (CrossKD). From the table, we see that the performance achieves the best when γ takes 0.5 across all action recognition models and image classification models on both UCF101 and CIFAR-100. This indicates that the distillation difficulty term and the DPP-based diversity term contribute equally when selecting those samples with both the low distillation difficulty and the high diversity. In another word, neither of the two terms should be more emphasized during the sample selection per epoch.

7 VISUALIZATION SUPPLEMENT

The following visualization results are obtained on UCF101 [12].

Sample distillation strength α **per epoch**. To illustrate the tendency of the sample distillation strength α across different epochs, we average the distillation strength scores of selected samples per epoch on UCF101 and depict the varying curves in Fig. 2. As shown in Fig. 2a, the distillation strength scores are less than 0.5 during the first 35 epochs, which suggests that the student learns more knowledge from the ground-truth labels than from teacher due to the existence of more difficult-to-transfer examples in the early stage. As the training progresses, the distillation strength score rises up. This is because those originally difficult-to-transfer samples may become easier to transfer and the student should learn more knowledge from the teacher. Similar observations are found in Fig. 2b, where the distillation strength scores are less than 0.5 during the first 15 epochs.

Selected sample number trend. To intuitively understand the dynamics of sample selection across different epochs, we draw the selected sample number bars in Fig. 3 on UCF101. This figure shows the number of selected samples in epoch 6, epoch 26, and epoch 46, respectively. As vividly shown in the bars, we observe that the number of selected samples varies across different classes, which reveals that some classes may be difficult to transfer while the others may be easy to transfer as the training progresses. The selected sample number in the early stage may be smaller than that

Table 10: Ablation studies on the sample selection interval with KD [6].

Interval	UCF101 [12]						CIFAR-100 [8]					
	Top1↑		Top5↑		Time(min)↓		Top1↑		Top5↑		Time(s)↓	
	Slowfast	VideoST	Slowfast	VideoST	Slowfast	VideoST	ResNet	WRN	ResNet	WRN	ResNet	WRN
1	**90.73**	**89.43**	**99.20**	**99.04**	4.6	11.1	**70.62**	**74.44**	**93.06**	**93.11**	40.6	33.6
3	90.56	87.89	98.90	98.13	2.0	5.2	70.58	74.28	92.98	93.02	30.2	26.3
5	89.88	87.71	98.80	98.72	1.5	3.9	70.55	74.23	92.85	92.79	26.6	22.8
8	89.22	87.23	98.41	98.34	1.2	3.1	70.24	74.04	92.56	92.61	25.8	22.5
10	89.09	86.88	97.98	98.05	**1.0**	**2.9**	70.09	73.86	92.17	92.38	**24.8**	**21.9**

Table 11: Ablation studies on the sample selection interval with CrossKD [14].

Interval	UCF101 [12]						CIFAR-100 [8]					
	Top1↑		Top5↑		Time(min)↓		Top1↑		Top5↑		Time(s)↓	
	Slowfast	VideoST	Slowfast	VideoST	Slowfast	VideoST	ResNet	WRN	ResNet	WRN	ResNet	WRN
1	**90.82**	**87.98**	98.91	**99.38**	4.9	11.6	**69.73**	**74.01**	93.09	**94.09**	41.6	36.2
3	90.56	87.03	98.90	98.53	1.9	5.0	69.42	73.92	**93.12**	93.97	31.9	26.5
5	90.04	86.94	**99.23**	97.16	1.5	3.8	69.21	73.87	92.93	93.87	26.8	23.6
8	89.34	86.72	98.03	96.91	1.1	3.0	69.04	73.52	92.76	93.52	25.3	22.9
10	88.92	86.02	97.51	96.82	**1.0**	**2.8**	68.93	73.08	92.31	93.29	**25.0**	**22.3**

Table 12: Ablation studies on the supervisor with KD [6].

Supervisor	UCF101 [12]				CIFAR-100 [8]			
	Top1 ↑		Top5 ↑		Top1 ↑		Top 5↑	
	SlowFast	VideoST	SlowFast	VideoST	ResNet	WRN	ResNet	WRN
GT+Teac.	**87.92**	**89.38**	**97.56**	**98.25**	**70.81**	**74.34**	**92.01**	**92.78**
Teacher	85.80	86.23	95.50	97.49	70.04	73.36	91.80	93.46
GT	85.15	86.17	95.80	97.03	69.06	73.26	91.12	92.03

Table 13: Ablation studies on the supervisor with CrossKD [14].

Supervisor	UCF101 [12]				CIFAR-100 [8]			
	Top1 ↑		Top5 ↑		Top1 ↑		Top5 ↑	
	SlowFast	VideoST	SlowFast	VideoST	ResNet	WRN	ResNet	WRN
GT+Teac.	**90.79**	**87.63**	**99.10**	**98.06**	**69.95**	**73.41**	**92.52**	**93.14**
Teacher	85.23	86.69	95.33	97.82	69.32	73.17	91.87	92.23
GT	85.15	86.17	95.80	97.03	69.06	73.26	91.12	92.03

Table 14: Ablations on the diversity parameter γ with KD [6].

γ	UCF101 [12]				CIFAR-100 [8]			
	Top1 ↑		Top5 ↑		Top1 ↑		Top5 ↑	
	SlowFast	VideoST	SlowFast	VideoST	ResNet	WRN	ResNet	WRN
0.0	87.74	86.42	97.85	97.51	68.55	72.84	91.32	92.65
0.1	88.08	86.71	98.02	97.73	68.71	73.38	91.56	92.83
0.3	88.93	87.13	98.33	98.16	69.28	74.01	92.18	93.30
0.5	**89.88**	**87.70**	**98.80**	**98.72**	**70.55**	**74.23**	**92.85**	**93.94**
0.7	89.03	87.26	98.47	98.43	69.29	74.07	92.09	93.58
0.9	88.52	86.74	98.17	97.97	68.47	73.72	91.74	93.03
1.0	88.18	86.52	97.98	97.85	68.03	73.50	91.52	92.80

in the later stage, and vice versa, which suggests that some difficult-to-transfer samples become easier to transfer with more epochs, which makes them be more easily to be selected. In addition, we show the images from the top six selected classes across different epochs in Fig. 4, which shows the selected classes vary from epoch 6 to epoch 26 and from epoch 26 to epoch 46. In another word, the sample distillation strength scores should also vary during student training.

Table 15: Ablation studies on the diversity parameter γ with CrossKD [14].

γ	UCF101 [12]				CIFAR-100 [8]			
	Top1 ↑		Top5 ↑		Top1 ↑		Top5 ↑	
	SlowFast	VideoST	SlowFast	VideoST	ResNet	WRN	ResNet	WRN
0.0	87.91	85.73	97.99	96.78	67.23	72.68	91.27	92.73
0.1	88.53	86.03	98.21	96.89	67.92	72.94	91.72	93.02
0.3	89.48	86.59	98.63	97.02	68.62	73.12	92.05	93.28
0.5	**90.04**	**86.94**	**99.23**	**97.16**	**69.21**	**73.87**	**92.53**	**93.87**
0.7	89.53	86.67	98.83	97.01	68.93	73.02	92.35	93.27
0.9	89.22	86.31	98.60	96.81	68.23	72.38	92.09	92.10
1.0	88.98	86.04	98.48	96.77	67.72	72.12	91.67	92.04

Moreover, we show how the number of selected samples changes in Fig. 5, which show the change number of the selected samples in most classes is between -5 and 5, i.e., reducing five samples or increasing 5 samples. We also show the images from the class in the top six change number list in Fig. 6. From the figure, we observe that there are two classes (i.e., "Biking" and "SalsaSpin") are the same, which indicates that the distillation difficulty changes are large. For example, the selected number from the class "Biking" is 9 in epoch 6, 17 in epoch 26, and 6 in epoch 46 (See Fig. 3), which reveals the fact that the samples in this class become easier to transfer in intermediate epochs, and there are no more knowledge to transfer as the training progresses which makes the select number be reduced in later epochs; the selected sample from the class "SalsaSpin" is 14 in epoch 6, 2 in epoch 26, and 10 in epoch 46 (See Fig. 3), which shows that the samples in this class are easy to transfer in the initial epochs, but some samples become difficult to transfer in intermediate epochs while some others become easier to transfer in later epochs since the student becomes stronger. This validates the fact that the samples plays different roles in knowledge distillation during different epochs, so we should treat them differently at sample level during student training for better performance.

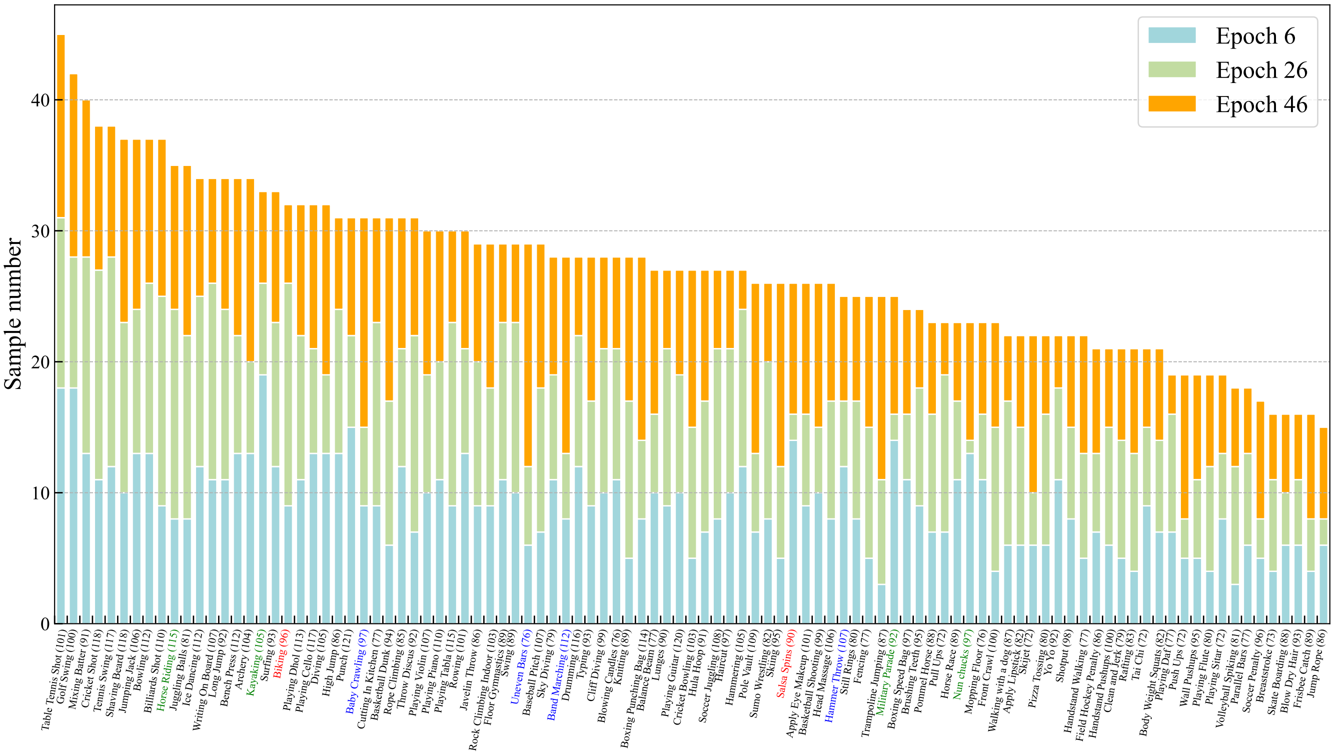
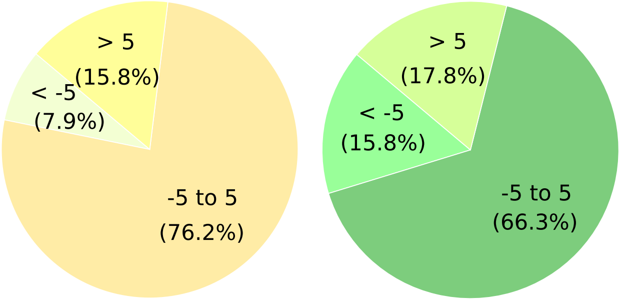


Figure 3: Trend of sample number per class from epoch 6 to 26, and from epoch 26 to 46. The digits in parenthesis denote sample number per class, the class name in green denotes it is in the top six sample change numbers per class from epoch 6 to 26, the class name in blue denotes that from epoch 26 to 46, and the class name in red denotes that from epoch 6 to 26 and from epoch 26 to 46.



(a) Selected sample number change from epoch 6 to 26.

(b) Selected sample number change from epoch 26 to 46.

Figure 5: Selected sample number change pie.

Figure 6: Top sample change number per class from epoch 6 to 26, and from epoch 26 to 46.

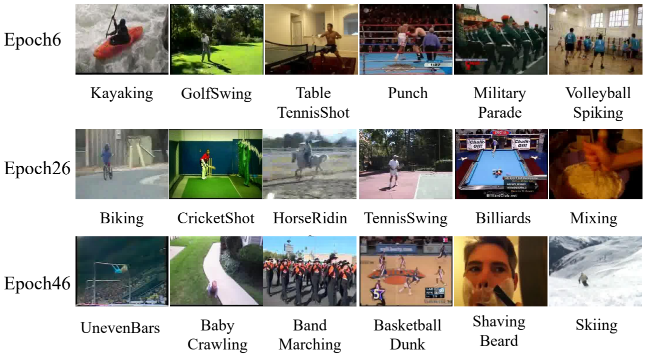


Figure 4: Top selected sample classes in epoch 6, epoch 26, and epoch 46.

REFERENCES

[1] Joo Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4724–4733.

[2] Jia Deng, Wei Dong, Richard Socher, Li Jia Li, Kai Li, and Li FeiFei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 248–255.

[3] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 6202–6211.

[4] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Frnd, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. 2017. The "something something" video database for learning and evaluating

visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 5843–5851.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[6] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint* arXiv:1503.02531 (2015).

[7] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[8] Alex Krizhevsky and Geoffrey Hinton. 2009. *Learning multiple layers of features from tiny images*. Technical Report. Canadian Institute for Advanced Research.

[9] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. 2023. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 1504–1512.

[10] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3202–3211.

[11] Cuong Pham, Van-Anh, Nguyen Trung, Le Dinh, and Phung Gustavo Carneiro. 2024. Frequency attention for knowledge distillation. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)* (2024), 2277–2286.

[12] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[13] Guiqin Wang, Peng Zhao, Yanjiang Shi, Cong Zhao, and Shusen Yang. 2024. Generative model-based feature knowledge distillation for action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 15474–15482.

[14] Jiabao Wang, Yuming Chen, Zhaohui Zheng, Xiang Li, Ming-Ming Cheng, and Qibin Hou. 2024. CrossKD: Cross-head knowledge distillation for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 16520–16530.

[15] Rui Wang, Yixue Hao, Long Hu, Xianzhi Li, Min Chen, Yiming Miao, and Iztok Humar. 2024. Efficient crowd counting via dual knowledge distillation. *IEEE Transactions on Image Processing (TIP)* 33 (2024), 569–583.

[16] Yi Xie, Huaidong Zhang, Xuemiao Xu, Jianqing Zhu, and Shengfeng He. 2023. Towards a smaller student: capacity dynamic distillation for efficient image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 16006–16015.

[17] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. 2020. Temporal pyramid network for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 591–600.

[18] Sergey Zagoruyko and Nikos Komodakis. 2016. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*. 19–22.

[19] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. 2022. Decoupled knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 11953–11962.

[20] Martin Zong, Zengyu Qiu, Xinzhu Ma, Kunlin Yang, Chunya Liu, Jun Hou, Shuai Yi, and Wanli Ouyang. 2023. Better teacher better student: dynamic prior knowledge for knowledge distillation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.