Use Your Head: Improving Long-Tail Video Recognition Supplementary Material

Dataset	Proposed Properties H%	F%	I	Cls.	Train	Val	Test
SSv2 [2]	26	0	79	174	168913	24777	N/A
SSv2-LT	9	32	500	174	50418	6960	2610
VideoLT [5]	23	0	43	1004	179334	25619	51239
VideoLT-LT	12	38	110	772	71207	7720	7720

Table 1. Original and curated (-LT) long-tail datasets.

1. Video-LT and SSv2-LT Datasets

In Section 3 in the main paper, we curated long-tail versions of SSv2 [2] and VideoLT [5]. More details are provided here.

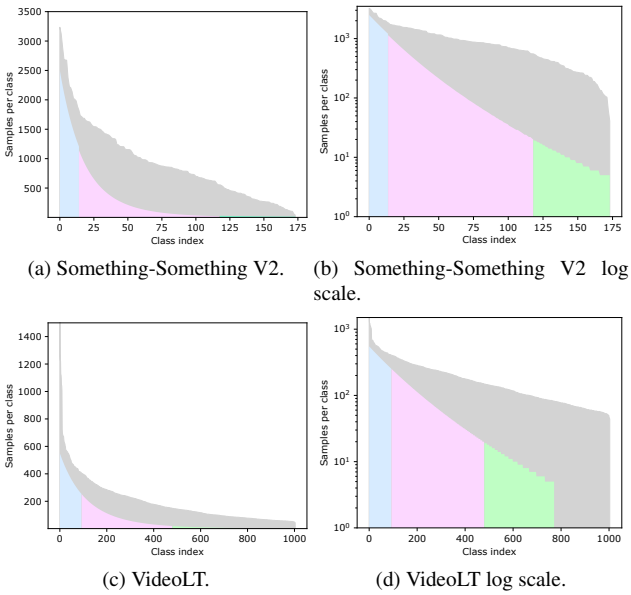
SSv2-LT: We follow the recipe used for ImageNet-LT and Places-LT from [3] and use the Pareto distribution with $\alpha = 6$ and a minimum class count of 5. We rank classes by their original size in the training set (*i.e*. the largest class in SSv2 is still the largest class in SSv2-LT and so on). We take a maximum class size of 2500, which is as large as it can be given the original and curated dataset sizes. Balanced training and validation sets are taken from the original training split, and the test set is taken from the original validation split (labels are not available for the test split from [2]).

VideoLT-LT: We use the same recipe as above, setting the maximum class size of 550, and keep the minimum as 5 and α as 6. We sample the proposed long-tail train split from the original VideoLT train split, and sample balanced val and test sets from the original unbalanced val and test test splits respectively. We do not include test videos with multiple labels (around 10%), and we do not include classes with fewer than 10 test samples. This maintains 772 classes, and ensures our smallest classes are evaluated robustly.

Class count distributions of the original datasets and the (-LT) curated versions are shown in regular and log scale in Fig. 1. Splits are shown in Table 1.

2. Motionformer parameters

Table 2 shows the parameter used for Motionformer [4] on EPIC-KITCHENS-100 and SSv2-LT. These are the defaults for EPIC-KITCHENS-100 [1] and Something-Something V2 [2] provided with the code for [4].



(a) Something-Something V2. (b) Something-Something V2 log scale.

(c) VideoLT. (d) VideoLT log scale.

Figure 1. Original datasets (grey) compared to our long-tail versions in standard scale (left) and log scale (right). SSv2 is top and VideoLT is bottom. Blue, pink and green regions show head, tail and few-shot classes in our proposed -LT splits.

References

[1] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling Egocentric Vision. *IJCV*, 2021. 1

[2] Raghav Goyal, Vincent Michalski, Joanna Materzy, Susanne Westphal, Heuna Kim, Valentin Haenel, Peter Yianilos, Moritz Mueller-freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. In *ICCV*, 2017. 1

[3] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019. 1

[4] Patrick Mandela, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F. Henriques. Keeping Your Eye on the Ball : Trajectory Attention in Video Transformers. In *NeurIPS*, 2021. 1, 2

	Parameter	Values
Model	Frame size	224x224
	Num frames	16
	Num blocks	12
	Num heads	12
	Embed dim	768
	Patch size	16
Train	Input augmentation	RandAugment
	Batch size	56
	Base lr	0.0001
	Momentum	0.9
	Weight decay	0.05
	Epochs	EPIC: 50, SSv2: 35
	Schedule gamma	0.1
	Schedule epochs	EPIC: 30,40, SSv2: 20, 30
	Optimiser	adamw
Test	Ensemble views	10
	Spatial crops	3

Table 2. Motionformer [4] parameters

[5] Xing Zhang, Zuxuan Wu, Zejia Weng, Huazhu Fu, Jingjing Chen, Yu-Gang Jiang, and Larry Davis. VideoLT: Large-scale Long-tailed Video Recognition. In *ICCV*, 2021. 1