Use Your Head: Improving Long-Tail Video Recognition

Toby Perrett Saptarshi Sinha Tilo Burghardt Majid Mirmehdi Dima Damen
<first>.<last>@bristol.ac.uk University of Bristol, UK

Abstract

This paper presents an investigation into long-tail video recognition. We demonstrate that, unlike naturally-collected video datasets and existing long-tail image benchmarks, current video benchmarks fall short on multiple long-tailed properties. Most critically, they lack few-shot classes in their tails. In response, we propose new video benchmarks that better assess long-tail recognition, by sampling subsets from two datasets: SSv2 and VideoLT.

We then propose a method, Long-Tail Mixed Reconstruction (LMR), which reduces overfitting to instances from few-shot classes by reconstructing them as weighted combinations of samples from head classes. LMR then employs label mixing to learn robust decision boundaries. It achieves state-of-the-art average class accuracy on EPIC-KITCHENS and the proposed SSv2-LT and VideoLT-LT. Benchmarks and code at: tobyperrett.github.io/lmr

1. Introduction

Advances in deep learning have been driven by increasing quantities of data to train larger and more sophisticated models. Landmark recognition datasets such as ImageNet [17] and Kinetics [10], amongst others, have fulfilled this need for data by first defining a taxonomy, and then scraping or crowd-sourcing until a sufficient number of examples are obtained for each class. They typically aim for balanced, or nearly balanced, class distributions. However, in practice, collecting enough examples for every object or action, including rare ones, remains challenging. Naturally occurring data is known to come from long-tail distributions, where it is often not possible to obtain a sufficient number of samples from classes in the tail.

In order to encourage methods to train effectively on long-tail data, image-recognition benchmarks include multiple naturally-collected[1] [24] as well as curated long-tail datasets [6, 9, 15, 37, 64]. In contrast, long-tail video recognition has been a less explored field. In Fig. 1, we compare image and video benchmarks, showcasing that none

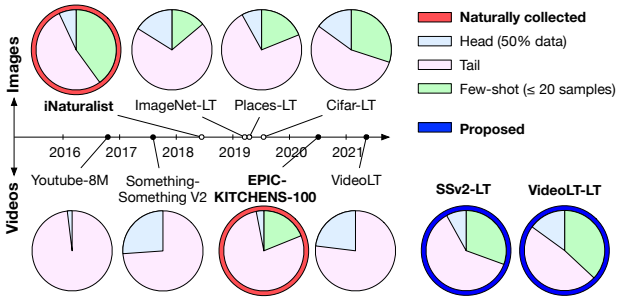


Figure 1. Long-tail image recognition datasets (top) [9, 37] aimed to curate similar distributions to the naturally-collected iNaturalist [24]. For video datasets (bottom), the naturally-collected EPIC-KITCHENS-100 [16] demonstrates a similar distribution of head/tail/few-shot classes to long-tail image datasets. In comparison, curated video datasets do not include *any* few-shot classes [1, 21, 71]. We propose two versions of existing datasets which do – SSv2-LT and VideoLT-LT. Head/tail/few-shot definitions in Fig. 2. Numeric comparison in Tab. 1.

of the curated video datasets to date contain any few-shot classes [1, 21, 71]. This is a critical oversight, as seminal research has highlighted that long-tail methods must *"learn accurate few-shot models for classes in the tail of the class distribution"* [64] and *"deal with imbalanced classification, few-shot learning"* [37]. In this paper, we follow the approach from [37] and re-sample videos to introduce long-tail versions of two video datasets.

We evaluate current long-tail recognition methods on our re-sampled long-tail video datasets and the naturally-collected EPIC-KITCHENS-100 dataset [16]. Unsurprisingly, when confronted with few-shot classes, current methods perform poorly due to a lack of sample diversity in the few-shot classes. We thus propose a new method that focuses on improving the performance on few-shot classes. Long-Tail Mixed Reconstruction (LMR) reconstructs few-shot samples as weighted combinations of head samples within the batch. A residual connection, weighted by the class size, is used to combine instances with their reconstructions. We use pairwise label mixing on these reconstructed samples to help learn robust class decision boundaries. Our key contributions are as follows:

- We compare image and video long-tail datasets, by providing a consistent definition of properties for long-tail

[1] We use the term 'naturally' to focus on the data collection. It does not imply footage of nature. We hope this footnote prevents any confusion.

class distributions.

- We curate new long-tail video benchmarks (-LT) which better test long-tail recognition performance.
- We propose a method, LMR, which increases the diversity in few-shot classes. It achieves highest average class accuracy across 3 benchmarks: naturally-collected EPIC-KITCHENS-100 and the two proposed curated benchmarks SSv2-LT and VideoLT-LT.

Sec. 2 reviews works which investigate long-tail characteristics, leading to the introduction of a set of properties and the comparison of existing long-tail benchmarks. Sec. 3 introduces new benchmarks and demonstrates experimentally the value of these long-tail properties. Sec. 4 summarises prior long-tail and few-shot video recognition approaches. Sec. 5 introduces LMR, our method for long-tail video recognition. Comparative analysis is given in Sec. 6. Finally, ablations on LMR are performed in Sec. 7.

2. Properties of Long-Tail Benchmarks

Established benchmarks for long-tail image recognition [37] have shaped the progress of long-tail methods. These followed earlier efforts that investigated the desired data distribution characteristics for long-tail benchmarks. In [6], experiments were performed with class counts that decay linearly or decay with a step-function. They noted that a larger imbalance between majority (now known as 'head') and minority (i.e. 'tail') classes increases difficulty and that a longer tail negatively affects classifier performance for both linear and step class count decays. Interestingly, imbalance was shown to affect higher complexity tasks (*e.g.* CIFAR) significantly more than lower complexity tasks (*e.g.* MNIST). Step and exponential class count decays were also investigated in [9], with similar conclusions. In [15], multiple long-tail versions of CIFAR [29] were curated by changing the minimum class size. Distribution characteristics were not explored numerically, but a drop in performance was reported as the number of samples per class decreased.

Despite the richness of these early findings, imbalance (*i.e.* the ratio between the largest and smallest class sizes) has become the primary metric for characterising long-tail benchmarks. However, imbalance ignores other critical characteristics such as the number of few-shot classes. To reflect this, we define three properties which together allow a more informed comparison of long-tail benchmarks. These are visualised in Fig. 2:

- **Head Length (H%):** The percentage of classes that formulates the majority of samples in the dataset. When classes are ranked by their size in the training set, these are the largest classes that together contribute $x\%$ of the training samples. While different values can be used for x, we follow prior work that used 50% of the data to represent head classes [4, 52]. We consider the head length as the ratio of head classes to all classes.
- **Few-Shot Length (F%):** The percentage of few-shot classes in the dataset, where a few shot class contains $\leq x$ training samples. Prior works use values between 5 and 50 for x [2, 8, 41, 46, 48, 58, 59, 61, 69, 75]. We follow long-tail image works and use 20 as the threshold for few-shot classes [37, 72].
- **Imbalance (I):** Previously used in [15], imbalance is the ratio between the size of the largest and smallest classes. Note that this metric alone does not provide a measure of how long-tailed a dataset is.

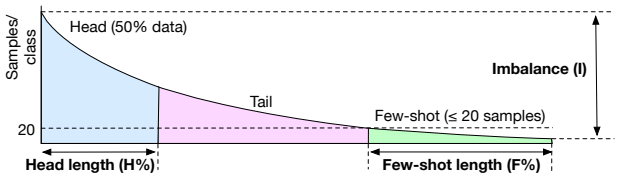


Figure 2. Visualisation of long-tail distribution properties: head length (H%), few-shot length (F%) and imbalance (I). Previous works have relied solely on imbalance, or used the terms "head", "mid" and "tail" to describe different parts of the distribution with arbitrarily chosen sizes. In this paper, we use consistent properties to compare long-tail benchmarks across images and videos.

These three properties are distribution agnostic, *i.e.* they can describe the properties of any benchmark whether the data is naturally-collected, or when it is sampled, no matter what distribution function is used. Using these three properties (H%, F%, I), we now quantitatively compare long-tail datasets across images and videos.

2.1. Long-Tail Image Datasets

The definitive example of a naturally-collected long-tail image recognition dataset is iNaturalist 2018 [24]. It is constructed from image and label contributions of plants and animals in the wild. As some species are rare, it would be very difficult to acquire more examples of these few-shot classes. As shown in Tab. 1, the iNaturalist image dataset has a head length of 7% (*i.e.* the 7% largest classes contribute 50% of the data), a few-shot length of 40% (*i.e.* 40% of the classes have 20 or fewer training examples) and an imbalance of 500. Thus, for methods to perform well on naturally-collected data, they must be good at learning a large number of few-shot classes.

Methods also evaluate on curated long-tail versions of large-scale datasets to avoid over-specialisation on iNaturalist. The widely used ImageNet-LT [37], Places-LT [37] and CIFAR-LT [15] re-sample from the original datasets and have comparable properties to the naturally-collected iNaturalist, making them suitable for evaluating methods that target long-tail recognition. As shown in Tab. 1, these

	Source	Dataset	Year	Proposed Properties H%	F%	I	Class size Max	Min	Num classes	Balanced test	Content
Images	Natural	iNaturalist [24]	2018	7	40	500	1000	2	8142	✓	Photos of species
	Resampled	ImageNet-LT [37]	2019	16	14	256	1280	5	1000	✓	Image recognition
	Resampled	Places-LT [37]	2019	8	19	996	4980	5	365	✓	Photos of scenes
	Resampled	Cifar-LT-100 [15]	2019	15	30	100	500	5	100	✓	Image recognition
Videos	Natural	EPIC-KITCHENS-100 Verbs [16]	2020	3	19	14848	14848	1	97	✗	Egocentric actions
	Collected	Youtube-8M [1]	2016	2	0	6409	788288	123	3862	✗	Youtube
	Collected	Something-Something V2 [21]	2017	26	0	79	3234	41	174	✗	Temporal reasoning
	Collected	VideoLT [71]	2021	23	0	43	1912	44	1004	✗	Youtube (fine-grained)
	Resampled	SSv2-LT (proposed)	2022	9	32	500	2500	5	174	✓	Temporal reasoning
	Resampled	VideoLT-LT (proposed)	2022	12	38	110	550	5	772	✓	Youtube (fine-grained)

Table 1. Comparison of datasets against long-tail properties: Head Length (H%), Few-Shot Length (F%) and Imbalance (I). Red highlighted rows contain naturally-collected datasets. The bottom two rows (blue) contain our proposed VideoLT-LT and SSv2-LT, which are curated to better match naturally-collected data than other video benchmarks.

have few-shot lengths of $14\%, 19\%$ and 30% respectively and head lengths of $\leq 16\%$.

2.2. Long-Tail Video Datasets

By analogy, one naturally-collected large-scale and long-tail video dataset is EPIC-KITCHENS-100 [16]. Collection was unscripted recording of several days of kitchen activities. The number of samples of an action class roughly correlates to the how frequently the action occurs in daily activities. Table 1 shows EPIC-KITCHENS-100 has a head length of 3% and a few-shot length of 19%.

There have been two attempts at curating video datasets to specifically test long-tail methods, Youtube8M [1] and VideoLT [71]. While these are appreciated efforts, they are far from ideal as long-tail benchmarks. Table 1 shows neither of these contain any few-shot classes (F% = 0), and VideoLT has a significantly smaller imbalance of 43 compared to $100 - 996$ for long-tail image datasets. We build on this effort to propose long-tail benchmarks that satisfy all the desired properties.

3. Proposed Long-Tail Video Benchmarks

Having identified weaknesses in current benchmarks used for long-tail video recognition, we first propose to use EPIC-KITCHENS-100 as it is naturally-collected and satisfies the long-tail properties (as defined in Sec. 2). We also propose to resample public video datasets, so their properties are in line with curated long-tail image datasets.

SSv2 [21] is chosen as it is widely considered to be a good test of temporal understanding and has previously been re-purposed for evaluating few-shot works [8, 77]. Similarly, VideoLT [71] targets fine-grained classes. We call these curated versions SSv2-LT and VideoLT-LT, and resample these following the recipe used in [37] for ImageNet-LT and Places-LT (sampling from a Pareto distribution with $\alpha = 6$). Table 1 demonstrates these curated versions match the desired properties as visualised in Fig. 1. For additional details including sampled number of instances per class, see Appendix A.

Before proceeding to the method, ablations are first performed at a dataset level, where different curated versions of SSv2-LT are compared to demonstrate the impact on long-tail properties and the effect of few-shot classes. Full implementation details of models and metrics will be given in Sec. 6, but for these ablations it suffices to say that Motionformer [39] is trained with cross-entropy, reporting average class accuracy over the test set, as well as over few-shot, tail and head classes.

3.1. Importance of Long-Tail Properties

In Sec. 2, we noted that prior works use Imbalance (I) to identify a dataset as being long-tailed [15, 50]. We quantitatively showcase that imbalance alone is insufficient by constructing four variants of SSv2-LT (A, B, C, D), with a fixed training set size = 50.4k and a fixed imbalance I = 500. We vary the head length H% and the few-shot length F% as shown in Fig. 3. Variant C (which uses an identical decay to ImageNet-LT and Places-LT [37]), highlighted in blue, is the version used throughout this paper and proposed as the long-tail benchmark SSv2-LT.

As H% decreases and F% increases (A $\rightarrow$ D), there are significant drops in few-shot, tail and overall accuracy (up to 9%), whereas head performance improves. This is indicative of the distribution becoming more long-tailed. Because this behaviour occurs with fixed I, it can be concluded that H% and F% are indeed necessary for comparison of long-tail distributions.

3.2. Effect of Few-Shot Classes

To showcase the importance of few shot classes, i.e. classes with $\leq$ 20 samples in training, we increment all classes in SSv2-LT with a fixed number of additional samples $+x$. We evaluate the performance over few-shot/tail/head classes[2] as we add $\{10, 20, 30, 40, 50\}$ samples per class. Fig. 4 shows that the accuracy on few-shot

[2]We maintain the set of classes in few-Shot/tail/head for direct comparison, i.e. even if the class has $>$ 20 samples after the addition.

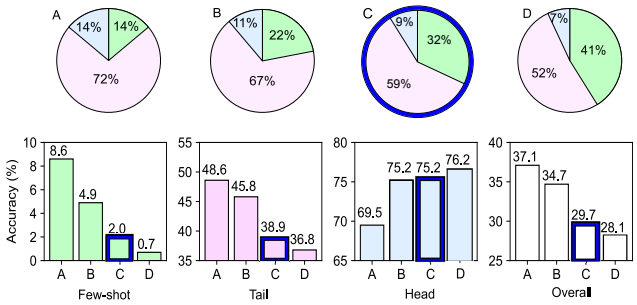


Figure 3. We compare four variants of SSv2-LT (A, B, C, D) with different H% and F% properties, while fixing I = 500, and the training dataset size = 50.4k. Top: percentage of head, tail and few-shot classes in each variant. Bottom: average class accuracy over the long-tail distribution. Variant C, highlighted in blue, is the proposed version used throughout the rest of this paper.

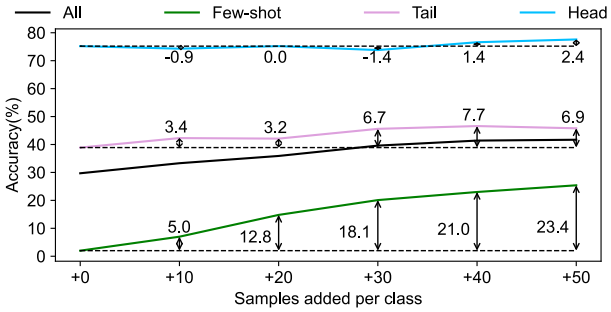


Figure 4. Effect of adding $+x$ samples per class on SSv2-LT. Average class accuracy is reported overall and for head, tail and few-shot classes. Per-case improvement reported next to arrow.

classes significantly increases when adding a small number of samples per class. The effect is smaller for tail classes and marginal for head. The maximum improvement of few-shot classes occurs around +20 samples/class, when no few-shot classes remain in training. To address this challenge, it is thus important to have a sufficient number of few-shot classes in long-tail benchmarks. Recall that naturally-collected datasets contain significant few-shot length (40% for iNaturalist and 19% for EPIC-KITCHENS-100).

4. Related Methods

Having justified our proposed benchmarks and before introducing our method, we first review long-tail and few-shot video recognition methods.

4.1. Long-Tail Methods

There are two main approaches to tackle long-tail recognition: re-weighting and re-balancing.

Re-weighting approaches impose a higher penalty when misclassifying samples from tail classes. This can be done by directly adjusting logits [32, 40, 47, 57, 66] or weighting the loss by class size [15, 53] or individual sample difficulty [18, 25, 31, 35, 43, 50]. Alternative approaches include label smoothing [73] and enforcing separation between class embeddings [27, 34, 49]. Re-weighting can also be achieved by enabling more experts to specialise on tail classes and combining predictions [7, 13, 30, 62, 72, 76].

Re-balancing approaches instead adjust the frequency at which examples from different classes are used in training, without adjusting the loss function. This can be achieved using a class-equalising feature bank [38] or more commonly by equal sampling from each class [22] or by instance difficulty [51, 67]. It has become standard practice to first train the representation with instance-balanced sampling [55] followed by class-balanced sampling [3, 26, 70].

Augmentations are known to introduce diversity into tail samples [33]. They combine the sample with a nearby class prototype [11], or create feature *clouds* to expand tail classes [36]. Further augmentation approaches include combining class-specific and class-generic features [12], using a separate classifier to identify head samples that can be adjusted and re-labeled as tail classes [28], or pasting tail foreground objects onto backgrounds from head classes [42]. Contrastive learning has also been used to improve representations [14, 74]. For video, Framestack [71] proposes temporally mixing up samples, frame-wise, based on average-precision during training.

Our proposed method, LMR, belongs to the re-balancing category. It is related to approaches for augmentation but differs in that it uses samples from *multiple* other classes, weighting the reconstruction by the class count and jointly reconstructing all samples in the batch.

4.2. Few-Shot Video Recognition

Despite the infancy of the long-tail video recognition field, the related field of few-shot video recognition has been more widely studied [5, 8, 20, 45, 56, 63, 65, 69, 75, 77, 78]. Instead of learning a long-tailed class distribution, few-shot methods learn to distinguish between a limited number of balanced few-shot classes (*e.g.* 5-way 5-shot). Few-shot video methods rely on attention between frames of the query video and all samples in the support set of each class [45, 63, 65, 78]. This requires the support set to be held in memory, which makes few-shot methods unsuitable for direct application to long-tail learning. Further, due to their design around balanced benchmarks, these methods cannot handle imbalance.

Our method takes inspiration from few-shot works in designing an approach for long-tail video recognition. In particular, image [19] and video [45] few-shot methods use a reconstruction technique to measure the similarity between a query and a class. A similar technique is used in [44] as input to a text captioning module. Each video is reconstructed from similar videos in the batch, using a cross-modal embedding space. In contrast to these works, we apply reconstruction *across classes* using multiple head samples to benefit those in the tail or those which are few shot. We also

make use of a residual connection to maintain knowledge of the sample itself. We detail our method next.

5. Method

When performing class-balanced sampling, instances from the tail are oversampled. This is particularly problematic for few-shot classes, where insufficient sample diversity results in overfitting. We propose Long-Tail Mixed Reconstruction (LMR), which aims to recover this diversity by computing a linear combination of the sample itself and weighted combinations of similar samples in the batch, weighted by the class size and followed by pairwise label mixing. In contrast to standard augmentation techniques, reconstructions are more representative of examples likely to be seen at test time, since they make use of visually similar samples from within the training set.

We first describe how classes are treated differently based on their count. We then proceed to describe our reconstruction and pairwise label mixing.

5.1. Long-Tailed Class Contribution

We consider the long-tailed class distribution of samples in the training set, and take C_y as the count of the class with label y. We define a contribution function $\mathbf{c}(y)$, per class, which we use later for reconstructing instances. We first calculate $\tilde{C}_y$ as the weight of class y:

$$\tilde{C}_y = \frac{1}{\log\left(C_y d + \epsilon\right)}, \tag{1}$$

where d controls the decay, and ϵ is a constant which ensures a positive denominator. These class weights can then be used to calculate the contribution function (low for head classes, high for tail):

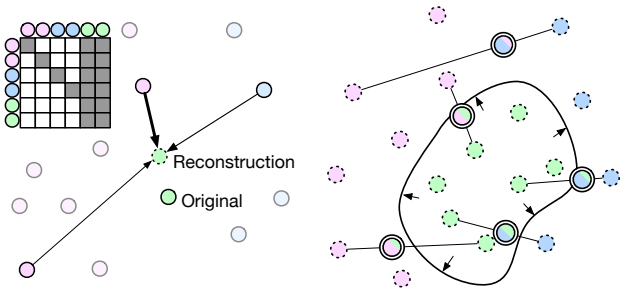
$$\mathbf{c}(y) = \frac{\tilde{C}_y - \min(\tilde{C}_y)}{\max(\tilde{C}_y) - \min(\tilde{C}_y)} l \,. \tag{2}$$

Here, $0 \leq l \leq 1$ is a hyperparameter controlling the contribution for the lowest class count. Note that these class contributions are established for the classes based on the training set, and not changed during training.

5.2. Long-Tail Mixed Reconstruction

Setup. Recognition methods combine a feature encoder $\mathbf{f}(\cdot)$ and a classifier $\mathbf{g}(\cdot)$. Data is fed to the model for training in the form of batches, where a batch X contains B videos $X = \{x_i : i = 1...B\}$ with associated labels $Y = \{y_i : i = 1...B\}$. Given the class contribution function from Eq. 2, we look up $\mathbf{c}(Y)$ for the samples in the batch, given their class labels.

To start, features for the batch are computed in the forward pass as $Z = \mathbf{f}(X)$. We propose a mixed reconstructor $\mathbf{mr}(\cdot, \cdot)$, which acts on features Z and labels Y, and returns a new reconstructed representation with an updated label for every video in the batch.



(a) The few-shot sample (green) is reconstructed as a weighted sum of other samples in the batch. Few-shot samples are prevented from contributing to the reconstructions through a mask (top left).

(b) Reconstructions (dotted outline) increase diversity for few-shot classes, expanding class boundaries. Robust boundaries are learned by pairwise mixing of reconstructions (double outline).

Figure 5. LMR overview: reconstruction (a) and label-mixing (b).

Sample reconstruction. We calculate cosine similarity $\mathbf{s}$ between all features within the batch, $S_{ij} = \mathbf{s}\left(Z_i, Z_j\right)$. Note that here, i denotes the feature to be reconstructed, and j denotes the feature being used for the reconstruction. We then calculate an exclusion mask E, avoiding self-weighting, i.e. samples should not contribute to their own reconstructions, and samples from few-shot classes are also avoided as these are already oversampled. The exclusion mask E is visualised in Fig. 5a, and calculated as:

$$E_{ij} = \begin{cases} 0 & \text{if } (i = j) \text{ or } (C_{y_i} \leq \omega) \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

where $\omega = 20$ is the few-shot threshold. Next, we apply a softmax operation over non-masked elements per row (*i.e.* one softmax per i), which calculates reconstruction weights W:

$$W_{ij} = \frac{\exp(S_{ij}) E_{ij}}{\sum_{k=1}^{B} \exp(S_{ik}) E_{ik}} \,. \tag{4}$$

We use a residual connection weighted by the class contribution – the smaller the class, the more the weighted features WZ contribute to the reconstruction of samples from that class. Specifically:

$$R = \mathbf{c}(Y) W Z + (1 - \mathbf{c}(Y)) Z \,, \tag{5}$$

where $\mathbf{c}(Y)$ (Eq. 2) are the contribution functions of the class labels in the batch and R are the reconstructed features. For few shot classes, the reconstruction is mostly formed from the weighted combination of other *similar* samples in the batch. Note that these reconstructions have the same class labels Y as the features Z they replace.

Pairwise label mixing. Once the reconstructions R are obtained, we take a step further by performing stochastic pairwise mixing (Fig. 5b). We use a mixing mask M such that:

$$M_{ij} = \begin{cases} \alpha_i & \text{if } (i = j) \\ 1 - \alpha_i & \text{if } (j = \beta_i) \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where α is a B-dimensional set of mixing weights, one for each sample. Following standard mixing, $\alpha_i = 1$ with probability 0.5, and randomly $0 \leq \alpha_i \leq 1$ otherwise. β is a B dimensional sample selector, that selects a different sample from the batch. $1 \leq \beta_i \leq B, \beta_i \neq i$ and $\beta_i \in \mathbb{N}$.

We apply the mixing mask M to our reconstructions R and their labels Y such that

$$\mathbf{mr}(Z, Y) = (MR, MY) \, . \quad (7)$$

We then pass these reconstructed and mixed features with the corresponding mixed labels to the classifier $\mathbf{g}$ to train.

5.3. Training and Inference

As customary [26], the classifier $\mathbf{g}$, acting on the backbone $\mathbf{f}$, is first pre-trained with instance-based sampling and cross-entropy. Afterwards, $\mathbf{g}$ is reset. LMR is then trained with class-balanced sampling and cross-entropy on $\mathbf{g}$. This is backpropagated through the mixed reconstructor $\mathbf{mr}$ and feature extractor $\mathbf{f}$. At inference, $\mathbf{mr}$ is discarded, as a suitable feature extractor $\mathbf{f}$ and classifier $\mathbf{g}$ have been learned for long-tail recognition. Each test sample/video is processed independently, *i.e*. there is no reconstruction, and labels and class counts are not used.

6. Experiments

We first perform comparative analysis on EPIC-KITCHENS-100, SSv2-LT and VideoLT-LT.

Metrics. The primary metric for long-tail video recognition is average class accuracy (Avg C/A), as it provides a fair evaluation when the test set is unbalanced. When the test set is balanced, as in the case of SSv2-LT and VideoLT-LT, Avg C/A and overall accuracy (Acc) are identical metrics. EPIC-KITCHENS-100 has an unbalanced test set so overall accuracy is also provided for reference. We also report average class accuracy for few-shot (marked "few" in tables), tail and head classes, as defined by the properties in Sec. 2.

Baselines. We compare against the following methods, also identified in [71] as suitable for long-tail video recognition:

- **CE:** Standard cross entropy trained with instance-balanced sampling.
- **EQL:** As in **CE**, but using an Equalization Loss [54], which reduces the penalty for misclassifying a head class as a tail class. This baseline is currently used by video transformer works to address class imbalance [65].
- **cRT:** Classifier Retraining [26]. This is now the standard practice of instance-balanced sampling, followed by a classifier reset and class-balanced sampling.
- **Mixup** [68]**:** Pairs of samples and their labels are mixed.
- **Framestack** [71]**:** Mixes up video frames based on a running total of class average precision.

Implementation Details. For all experiments on EPIC-KITCHENS-100 and SSv2-LT, we use Motionformer [39], a spatio-temporal transformer with attention guided by trajectories which achieves strong results on EPIC-KITCHENS-100 and SSv2. We use the default configuration of 16 frame input and 224×224 resolution with 16×16 patches. We train on 8×V100 GPUs, with a distributed batch of 56 samples. To enable processing on multiple GPUs, we maintain a feature bank of previous iterations per GPU. Other details (architecture, optimisation *etc*.) follow the default code of Motionformer and are noted in Appendix B. For all methods apart from CE and EQL, we follow the cRT disentanglement approach [26]. We first train end-to-end using instance-balanced sampling with a cross-entropy loss. We then reset the classifier and switch to class-balanced sampling for a full training run.

For VideoLT-LT experiments, we use the codebase provided with the original dataset and accompanying method Framestack [71] to be directly comparable to prior works. It uses pre-extracted ResNet-50 [23] frame features with a non-linear classifier and score aggregation. We use the default batch size of 128 samples trained on 1×P100 GPU.

For LMR, the few-shot threshold is $\omega = 20$. Decay and scaling parameters for the contribution function are $d = 0.25$ and $l = 0.6$ for SSv2-LT and VideoLT-LT, and $d = 0.15$ and $l = 1.0$ for EPIC-KITCHENS-100 as it has a smaller minimum class size.

Results. Table 2 shows the results for EPIC-KITCHENS-100, SSv2-LT and VideoLT-LT. LMR performs best on all datasets for average class accuracy. Note that prior results were reported on datasets that did not contain any few shot classes (see Sec 2.1). By evaluating on EPIC-KITCHENS-100, and proposing benchmarks with few-shot classes, we can expose the limitations of these methods previously deemed competitive for long-tail video recognition. LMR also obtains the best results on few-shot classes (highlighted in green) on all datasets. For tail classes, LMR performs comparably or outperforms prior baselines. For head classes, LMR performs comparably to long-tail baselines on EPIC-KITCHENS-100 and SSv2-LT, but takes a bigger hit on VideoLT-LT. We do not change any of the hyperparameters across datasets for fairer comparison, but consider results can be further improved if optimised per dataset.

Figure 6 shows class improvements of LMR compared to CE on EPIC-KITCHENS-100. Significant improvements are seen on smaller classes (few-shot and end of tail). Some head classes drop in performance, particularly the largest. Similar trends were found on SSv2-LT and VideoLT-LT.

Figure 7 shows selected examples from all datasets. CE tends to predict few-shot classes as visually similar head

Method	EPIC-KITCHENS-100 Few	Tail	Head	Avg C/A	Acc	SSv2-LT Few	Tail	Head	Avg C/A = Acc	VideoLT-LT Few	Tail	Head	Avg C/A = Acc
CE	0.0	12.3	**55.2**	21.2	**63.5**	2.0	38.9	**75.2**	29.7	17.4	51.1	**75.9**	41.0
EQL [54]	0.0	12.4	55.0	21.1	63.3	3.1	39.0	**75.2**	30.1	17.4	51.0	75.4	40.9
cRT [26]	21.4	35.0	51.1	36.9	50.1	14.9	45.6	58.6	36.5	30.5	**56.9**	64.0	47.5
Mixup [68]	25.8	33.8	51.7	36.8	51.7	17.4	**46.6**	57.1	37.8	15.8	48.9	72.5	38.9
Framestack [71]	23.0	33.6	52.1	36.5	52.5	15.5	46.1	61.9	37.2	18.2	51.8	74.5	41.5
LMR	**35.7**	**36.8**	51.1	**39.7**	51.3	**17.9**	46.5	61.0	**38.3**	**34.8**	56.8	62.1	**48.9**

Table 2. Long-tail results on EPIC-KITCHENS-100 Verbs Val set [16], SSv2-LT and VideoLT-LT. Note that average class accuracy (Avg C/A) is the same as overall accuracy (Acc) for balanced test sets (SSv2-LT and VideoLT-LT). EPIC-KITCHENS-100 has an unbalanced test set, so overall accuracy, which favours over-prediction of head classes, is provided for reference. LMR obtains the highest average class accuracy on all datasets, as well as the highest average class accuracy over few-shot classes.

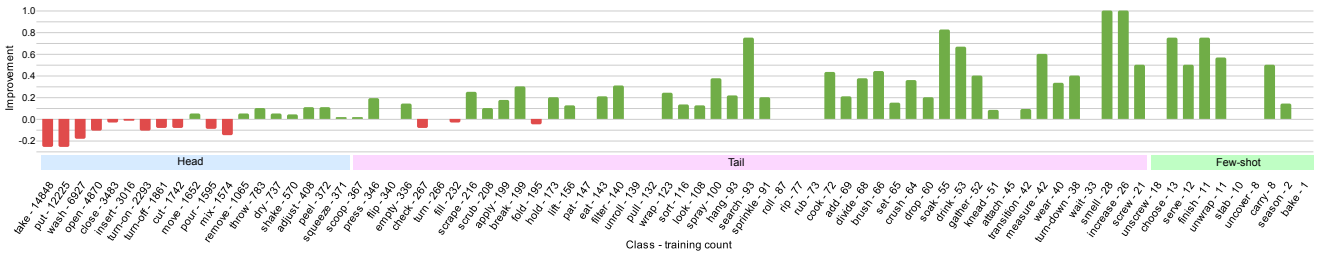


Figure 6. Improvements of LMR over CE on EPIC-KITCHENS-100. Classes are ordered by size and marked as head/tail/few-shot.

classes. For example, on EPIC-KITCHENS-100, CE misclassifies the few-shot "carry" as the head class "put" due to visual similarity of holding the cup. Consistently, LMR predicts the few-shot class correctly. A failure case is shown for SSv2-LT, where LMR predicts the head class "throwing something" as the tail class "throwing something in the air and letting it fall."

7. Ablations

We perform all ablations on EPIC-KITCHENS-100 and SSv2-LT using the Motionformer backbone.

LMR Ablation. Table 3 ablates the design choices of LMR against the full version (first row). First, class contributions are replaced by a constant (0.5 in A and 1 in B). When reconstructions are used solely, without the residual connection (B), performance decreases dramatically. Using label mixing without reconstructions is shown in (C) as well as reconstructions without label mixing (D). Interestingly, label mixing has a bigger impact on performance for SSv2-LT than EPIC-KITCHENS-100.

Contribution parameters. Reconstructions are combined with original representations according to the contribution function $\mathbf{c}(\mathbf{Y})$ in Eq. 5, which maps class count to a contribution between 0 and 1. It is parameterised by the decay d and the contribution l for the lowest class count. First, d is fixed at 0.25 and l is varied between 0.0 and 1.0. Results are shown in Tab. 4, where 0.6 performs best on the few-shot classes and overall. Next, l is fixed at 0.6 and d is varied, with results shown in Tab. 5. In both cases, results have a region of stability, with the best combination being $l = 0.6$ and $d = 0.25$.

Number of Samples Used for Reconstruction. We assess

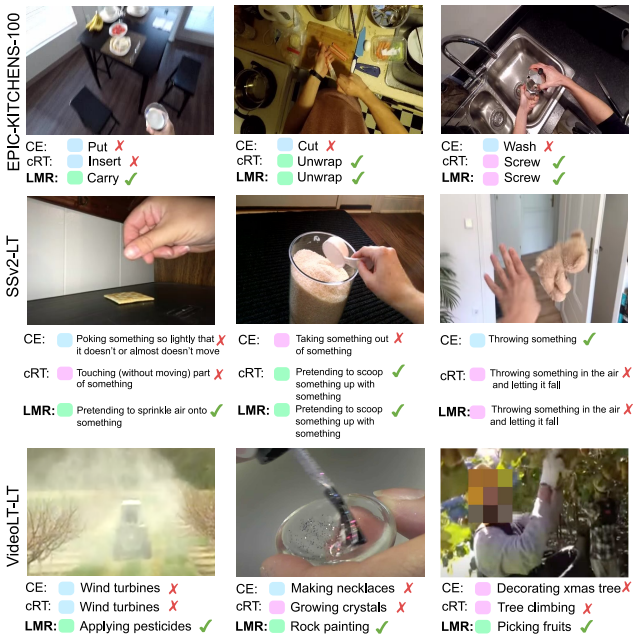


Figure 7. Qualitative examples from all benchmarks comparing CE, cRT and the proposed LMR. Blue, pink and green indicate whether the prediction is from a head, tail or few-shot class.

the impact of the number of samples in the batch used in the reconstruction process (B). Table 6 shows how varying the number of samples affects overall performance on SSv2-LT. Best performance is reported at our default of 56 samples.

Threshold for Masked Classes in Reconstruction. The threshold ω, used for masking in Eq. 3, is by default set to 20, which is the threshold for few-shot classes. The masking is used to prevent few-shot samples contributing to

Method Variant	EPIC-KITCHENS-100 Few	Tail	Head	Avg C/A	SSv2-LT Few	Tail	Head	Avg C/A
LMR	**35.7**	36.8	51.1	**39.7**	17.9	**46.5**	61.0	**38.3**
(A) Constant contribution [replace Eq. 2 with $c(y) = 0.5$]	34.1	**37.1**	49.3	39.3	16.8	44.9	**61.9**	37.1
(B) No original representation in reconstruction [replace Eq. 2 with $c(y) = 1$]	4.5	2.5	5.0	3.4	4.8	7.4	18.1	6.0
(C) No reconstruction [replace Eq. 5 with $R = Z$]	20.2	36.7	52.0	38.1	16.7	46.4	59.4	37.6
(D) No pairwise label mixing [replace Eq. 7 with $\mathbf{mr}(Z, Y) = (R, Y)$]	24.6	33.9	**53.2**	37.1	**18.0**	45.9	59.0	37.9

Table 3. Ablating LMR on EPIC-KITCHENS-100 and SSv2-LT.

l	Few	Tail	Head	Acc
0.0	16.7	46.4	59.4	37.6
0.2	16.9	46.3	59.0	37.7
0.4	16.9	46.0	58.6	37.6
0.6	**17.9**	46.5	61.0	**38.3**
0.8	17.6	46.5	**61.9**	38.2
1.0	16.2	**46.7**	60.5	37.7

Table 4. Effect of changing l, the contribution applied to the lowest class count on SSv2-LT. A higher l means reconstructions contribute more to the representations.

d	Few	Tail	Head	Acc
0.11	17.7	46.1	60.0	38.0
0.25	**17.9**	46.5	**61.0**	**38.3**
0.5	13.7	47.5	60.0	37.5
1.0	12.4	**48.0**	59.5	37.4

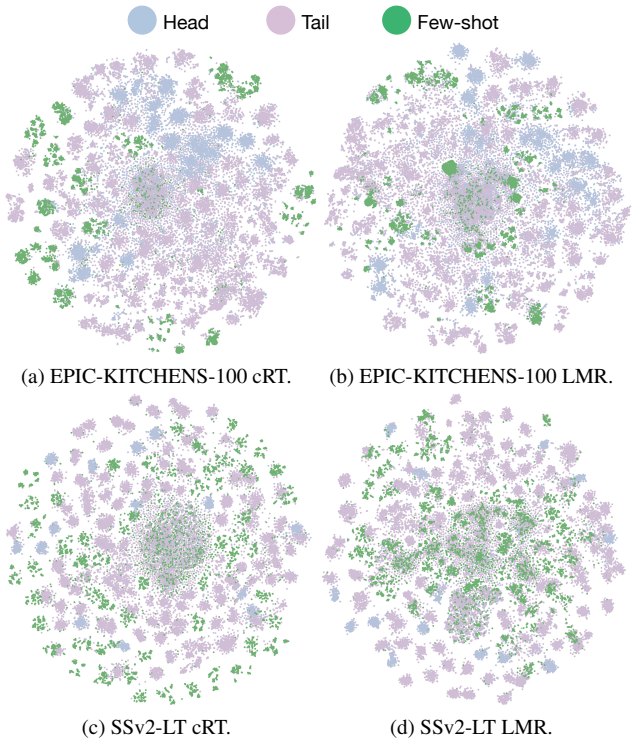
Table 5. Effect of changing d, the decay of the class count contribution, on SSv2-LT. A lower d means the contributions of reconstructions decay faster as class counts increase.

B	Few	Tail	Head	Acc
14	17.1	**46.7**	60.0	38.0
56	**17.9**	46.5	**61.0**	**38.3**
224	17.6	46.2	60.0	38.0
896	17.6	46.0	59.0	37.9

Table 6. Effect of varying the number of samples (B) used for reconstruction on SSv2-LT.

ω	Few	Tail	Head	Acc
0	**18.0**	45.9	59.5	37.9
20	17.9	**46.5**	**61.0**	**38.3**
50	17.6	46.2	59.5	37.9
500	17.5	46.2	59.0	37.9

Table 7. Effect of changing ω on SSv2-LT, which is the minimum class size threshold for the reconstruction mask.



(a) EPIC-KITCHENS-100 cRT. (b) EPIC-KITCHENS-100 LMR.

(c) SSv2-LT cRT. (d) SSv2-LT LMR.

Figure 8. Effect of LMR on EPIC-KITCHENS-100 (top) and SSv2-LT (bottom) t-SNE projections. Without reconstruction (left), samples from few-shot classes (green) are pushed to the edge, and tightly clustered. With LMR (right), the few-shot clusters are larger and closer to the centre, i.e. in closer proximity to head (blue) and tail (pink) classes. This gives more robust boundaries as they are bordering more classes.

the reconstruction of other samples. Table 7 shows the effect of varying ω. Best performance is obtained at $\omega = 20$.

Visualising LMR . Fig. 8 shows t-SNE [60] projections of representations without LMR (*i.e.* cRT) and with. cRT pushes the few shot classes (green) to the periphery. LMR results in larger, *i.e.* more diverse, few-shot clusters towards the centre of the projection. This indicates a higher proximity to head and tail classes which creates robust class boundaries and better generality to unseen test samples.

8. Conclusion

In this paper, we defined a set of properties, enabling quantitative comparison of long-tail distributions. We showcased that curated long-tail image datasets are comparable to naturally-collected ones, while previously proposed video datasets fall short. Based on these findings, we proposed new benchmarks, SSv2-LT and VideoLT-LT, and suggested their use, alongside EPIC-KITCHENS-100, for evaluating long-tail video recognition.

We proposed LMR, a method for long-tail video recognition, which reconstructs few-shot samples as weighted combinations of other samples in the batch. A residual connection, weighted by the class size, combines instances with their reconstructions, followed by pairwise label mixing. LMR reduces overfitting to instances from few-shot classes, and outperforms prior methods on the three benchmarks.

We hope our proposed benchmarks and method will provide a foundation for long-tail video recognition, and encourage further contributions applicable to naturally-collected data.

Acknowledgments. We use publicly available datasets and publish our proposed benchmarks. Research is funded by EPSRC UMPIRE (EP/T004991/1), EPSRC SPHERE Next Steps (EP/R005273/1), EPSRC DTP and EPSRC PG Visual AI (EP/T028572/1). We acknowledge the use of HPC Tier 2 Facility Jade 2 and Bristol's Blue Crystal 4 facility.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. YouTube-8M: A Large-Scale Video Classification Benchmark. In *arXiv*, 2016. 1, 3

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Nueral Information Processing Systems*, 2022. 2

[3] Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, and Shu Kong. Long-tailed recognition via weight balancing. In *CVPR*, 2022. 4

[4] Chris Anderson. *The Long Tail: Why the Future of Business Is Selling Less of More*. Hachette Books, 2006. 2

[5] Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. TARN: Temporal Attentive Relation Network for Few-Shot and Zero-Shot Action Recognition. In *BMVC*, 2019. 4

[6] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, (106):249–259, 2018. 1, 2

[7] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. ACE: Ally Complementary Experts for Solving Long-Tailed Recognition in One-Shot. In *ICCV*, 2021. 4

[8] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-Shot Video Classification via Temporal Alignment. In *CVPR*, 2020. 2, 3, 4

[9] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *NeurIPS*, 2019. 1, 2

[10] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *CVPR*, 2017. 1

[11] Xiaohua Chen, Yucan Zhou, Dayan Wu, Wanqian Zhang, Yu Zhou, Bo Li, and Weiping Wang. Imagine by reasoning: A reasoning-based implicit semantic data augmentation for long-tailed classification. In *AAAI*, 2022. 4

[12] Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling. Feature Space Augmentation for Long-Tailed Data. In *ECCV*, 2020. 4

[13] Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. Reslt: Residual learning for long-tailed recognition. *IEEE TPAMI*, 2022. 4

[14] Jiequan Cui, Zhisheng Zhong, Shu Liu, Yu Bei, and Jia Jiaya. Parametric Contrastive Learning. In *ICCV*, 2021. 4

[15] Yin Cui, Menglin Jia, Tsung Yi Lin, Yang Song, and Serge Belongie. Class-Balanced Loss Based on Effective Number of Samples. *CVPR*, 2019. 1, 2, 3, 4

[16] 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, 3, 7, 11

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 1

[18] Zongyong Deng, Hao Liu, Yaoxing Wang, Chenyang Wang, Zekuan Yu, and Xuehong Sun. PML : Progressive Margin Loss for Long-tailed Age Classification. In *CVPR*, 2021. 4

[19] Carl Doersch, Ankush Gupta, and Andrew Zisserman. CrossTransformers: Spatially-Aware Few-Shot Transfer. In *NeurIPS*, 2020. 4

[20] Sai Kumar Dwivedi, Vikram Gupta, Rahul Mitra, Shuaib Ahmed, and Arjun Jain. ProtoGAN: Towards Few Shot Learning for Action Recognition. In *CVPR*, 2019. 4

[21] 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, 11

[22] Agrim Gupta, Piotr Dollár, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 4

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 6

[24] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist Species Classification and Detection Dataset. In *CVPR*, 2018. 1, 2, 3

[25] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *CVPR*, 2020. 4

[26] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling Representation and Classifier for Long-Tailed Recognition. In *ICLR*, 2020. 4, 6, 7

[27] Tejaswi Kasarla, Gertjan J. Burghouts, Max van Spengler, Elise van der Pol, Rita Cucchiara, and Pascal Mettes. Maximum Class Separation as Inductive Bias in One Matrix. In *NeurIPS*, 2022. 4

[28] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced Classification via Major-to-minor Translation. In *CVPR*, 2020. 4

[29] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. 2009. 2

[30] Bolian Li, Zongbo Han, Haining Li, Huazhu Fu, and Changqing Zhang. Trustworthy long-tailed classification. In *CVPR*, 2022. 4

[31] Bo Li, Yongqiang Yao, Jingru Tan, Gang Zhang, Fengwei Yu, Jianwei Lu, and Ye Luo. Equalized focal loss for dense long-tailed object detection. In *CVPR*, 2022. 4

[32] Mengke Li, Yiu-ming Cheung, and Yang Lu. Long-tailed visual recognition via gaussian clouded logit adjustment. In *CVPR*, 2022. 4

[33] Shuang Li, Kaixiong Gong, Chi Harold, Liu Yulin, Wang Feng, and Qiao Xinjing. MetaSAug : Meta Semantic Augmentation for Long-Tailed Visual Recognition. In *CVPR*, 2021. 4

[34] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S. Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *CVPR*, 2022. 4

[35] Tsung-yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. In *ICCV*, 2017. 4

[36] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *CVPR*, 2020. 4

[37] 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, 2, 3, 11

[38] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Open long-tailed recognition in a dynamic world. *IEEE TPAMI*, 2022. 4

[39] 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. 3, 6, 11, 12

[40] Aditya Krishna Menon, Sadeep Jayanumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kuma. Long-Tail Learning via Logit Adjustment. In *ICLR*, 2021. 4

[41] Ashish Mishra, Vinay Kumar Verma, M Shiva Krishna Reddy, S Arulkumar, Piyush Rai, and Anurag Mittal. A Generative Approach to Zero-Shot and Few-Shot Action Recognition. In *WACV*, 2018. 2

[42] S. Park, Y. Hong, B. Heo, S. Yun, and J. Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *CVPR*, 2022. 4

[43] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. Influence-Balanced Loss for Imbalanced Visual Classification. In *ICCV*, 2021. 4

[44] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, João Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *ICLR*, 2021. 4

[45] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-Relational CrossTransformers for Few-Shot Action Recognition. In *CVPR*, 2021. 4

[46] Sachin Ravi and Hugo Larochelle. Optimizaion as a Model for Few-Shot Learning. In *ICLR*, 2017. 2

[47] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced Meta-Softmax for Long-Tailed Visual Recognition. In *NeurIPS*, 2020. 4

[48] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-Learning for Semi-Supervised Few-Shot Classification. In *ICLR*, 2018. 2

[49] Dvir Samuel and Gal Chechik. Distributional Robustness Loss for Long-tail Learning. In *ICCV*, 2021. 4

[50] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *NeurIPS*, 2019. 3, 4

[51] Saptarshi Sinha, Hiroki Ohashi, and Katsuyuki Nakamura. Class-Difficulty Based Methods for Long-Tailed Visual Recognition. *IJCV*, 2022. 4

[52] Susan Starr and Jeff Williams. The Long Tail: a Usage Analysis of Pre-1993 Print Biomedical Journal Literature. *Journal of the Medical Library Association*, 96(1), 2008. 2

[53] Jingru Tan, Xin Lu, Gang Zhang, Changqing Yin, and Quanquan Li. Equalization loss v2: A new gradient balance approach for long-tailed object detection. In *CVPR*, 2021. 4

[54] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *CVPR*, 2020. 6, 7

[55] Kaihua Tang, Mingyuan Tao, Jiaxin Qi, Zhenguang Liu, and Hanwang Zhang. Invariant feature learning for generalized long-tailed classification. In *ECCV*, 2022. 4

[56] Anirudh Thatipelli, Sanath Narayan, Salman Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Bernard Ghanem. Spatio-Temporal Relation Modeling for Few-shot Action Recognition. In *CVPR*, 2022. 4

[57] Junjiao Tian, Yen-Cheng Liu, Nathaniel Glaser, Yen-Chang Hsy, and Zsolt Kira. Posterior Re-calibration for Imbalanced Datasets. In *NeurIPS*, 2020. 4

[58] Pavel Tokmakov, Yu-Xiong Wang, and Martial Hebert. Learning Compositional Representations for Few-Shot Recognition. In *ICCV*, 2019. 2

[59] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples. In *International Conference on Learning Representations*, 2020. 2

[60] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[61] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching Networks for One Shot Learning. In *NeurIPS*, 2016. 2

[62] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *ICLR*, 2021. 4

[63] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Mingqian Tang, Zhengrong Zuo, Changxin Gao, Rong Jin, and Nong Sang. Hybrid relation guided set matching for few-shot action recognition. In *CVPR*, 2022. 4

[64] Yu Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *NeurIPS*, 2017. 1

[65] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. MeMViT: Memory-Augmented Multiscale Vision Transformer for Efficient Long-Term Video Recognition. In *CVPR*, 2022. 4, 6

[66] Tong Wu, Ziwei Liu, Qingqiu Huang, Yu Wang, and Dahua Lin. Adversarial robustness under long-tailed distribution. In *CVPR*, 2021. 4

[67] Sihao Yu, Jiafeng Guo, Ruqing Zhang, Yixing Fan, Zizhen Wang, and Xueqi Cheng. A re-balancing strategy for class-imbalanced classification based on instance difficulty. In *CVPR*, 2022. 4

[68] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup : Beyond Empirical Risk Minimization. In *ICLR*, 2018. 6, 7

[69] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip H S Torr, and Piotr Koniusz. Few-shot Action Recognition with Permutation-invariant Attention. In *ECCV*, 2020. 2, 4

[70] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *CVPR*, 2021. 4

[71] 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, 3, 4, 6, 7, 11

[72] Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Test-Agnostic Long-Tailed Recognition by Test-Time Aggregating Diverse Experts with Self-Supervision. In *Advances in Nerual Information Processing Systems Workshop*, 2021. 2, 4

[73] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *CVPR*, 2021. 4

[74] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *CVPR*, 2022. 4

[75] Linchao Zhu and Yi Yang. Compound Memory Networks for Few-Shot Video Classification. In *ECCV*, 2018. 2, 4

[76] Linchao Zhu and Yi Yang. Inflated episodic memory with region self-attention for long-tailed visual recognition. In *CVPR*, 2020. 4

[77] Linchao Zhu and Yi Yang. Label Independent Memory for Semi-Supervised Few-shot Video Classification. *IEEE TPAMI*, 14(8), 2020. 3, 4

[78] Xiatian Zhu, Antoine Toisoul, Juan-Manuel Pérez-Rúa, Li Zhang, Brais Martinez, and Tao Xiang. Few-shot Action Recognition with Prototype-centered Attentive Learning. In *BMVC*, 2021. 4

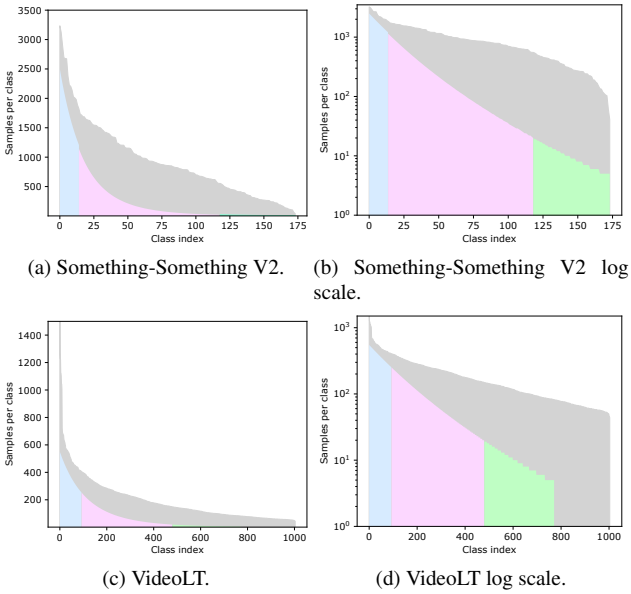
A. Video-LT and SSv2-LT Datasets

In Section 3, we curated long-tail versions of SSv2 [21] and VideoLT [71]. More details are provided here.

SSv2-LT: We follow the recipe used for ImageNet-LT and Places-LT from [37] 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 [21]).

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. 9. Splits are shown in Table 8.



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

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

Figure 9. 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.

Dataset	Proposed Properties H%	F%	I	Cls.	Train	Val	Test
SSv2 [21]	26	0	79	174	168913	24777	N/A
SSv2-LT	9	32	500	174	50418	6960	2610
VideoLT [71]	23	0	43	1004	179334	25619	51239
VideoLT-LT	12	38	110	772	71207	7720	7720

Table 8. Original and curated (-LT) long-tail datasets.

B. Motionformer parameters

Table 9 shows the parameter used for Motionformer [39] on EPIC-KITCHENS-100 and SSv2-LT. These are the defaults for EPIC-KITCHENS-100 [16] and Something-Something V2 [21] provided with the code for [39].

	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 instance bal/0.00001 class bal
	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 9. Motionformer [39] parameters