MEID: Mixture-of-Experts with Internal Distillation for Long-Tailed Video Recognition

Xinjie Li, Huijuan Xu

Pennsylvania State University, University Park, USA
{xql5497, hkx5063}@psu.edu

Abstract

The long-tailed video recognition problem is especially challenging, as videos tend to be long and untrimmed, and each video may contain multiple classes, causing frame-level class imbalance. The previous method tackles the long-tailed video recognition only through frame-level sampling for class re-balance without distinguishing the frame-level feature representation between head and tail classes. To improve the frame-level feature representation of tail classes, we modulate the frame-level features with an auxiliary distillation loss to reduce the distribution distance between head and tail classes. Moreover, we design a mixture-of-experts framework with two different expert designs, i.e., the first expert with an attention-based classification network handling the original long-tailed distribution, and the second expert dealing with the re-balanced distribution from class-balanced sampling. Notably, in the second expert, we specifically focus on the frames unsolved by the first expert by designing a complementary frame selection module, which inherits the attention weights from the first expert and selects frames with low attention weights, and we also enhance the motion feature representation for these selected frames. To highlight the multi-label challenge in long-tailed video recognition, we create two additional benchmarks based on Charades and CharadesEgo videos with the multi-label property, called CharadesLT and CharadesEgoLT. Extensive experiments are conducted on the existing long-tailed video benchmark VideoLT and the two new benchmarks to verify the effectiveness of our proposed method with state-of-the-art performance. The code and proposed benchmarks are released at https://github.com/VisionLanguageLab/MEID.

Introduction

Deep neural networks have achieved significant success in the video classification task (Lin, Gan, and Han 2019; Feichtenhofer et al. 2019) due to the networks' data-driven nature and the availability of large-scale video data. However, imbalanced data is ubiquitous in the real world, and model training on the long-tailed distribution may cause the under-representation of samples from tail classes. Furthermore, as shown in Fig. 1, long and untrimmed videos may contain multiple labels simultaneously, causing frame-level class imbalance, which makes long-tailed video recognition more challenging.

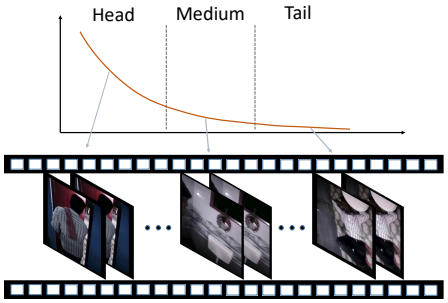

Figure 1: One example video contains multiple labels, causing frame-level class imbalance. Labels from left to right: opening a door (head class), watching something/someone/themselves in a mirror (medium class) and lying on the floor (tail class). The illustration video is from our proposed CharadesLT benchmark.

Recently, Zhang et al. (Zhang et al. 2021b) release the VideoLT dataset to facilitate the research on long-tailed video recognition, and propose the FrameStack model, which performs sampling at the frame level using the running average precision as guidance. While FrameStack (Zhang et al. 2021b) emphasizes the importance of frame-level class re-balance, it does not distinguish the frame-level feature representation between head and tail classes. Another work tackling the long-tailed video recognition task is MEDC (Hu, Gao, and Xu 2022) which models both the intra-class distribution and the inter-class distribution for long-tailed classes. However, MEDC (Hu, Gao, and Xu 2022) focuses on calibrating the intra-class distribution to adapt to different inter-class distributions in the test phase instead of considering the multi-label frame-level imbalance.

Besides these preliminary explorations in the long-tailed video recognition task, there already exists a lot of work in the long-tailed image recognition task and the common characteristic is increasing the impact of tail classes. For example, re-weighting based methods (Cui et al. 2019; Cao et al. 2019; Tan et al. 2020) try to find a better loss weight for the tail classes to boost the performance of tail classes, and

re-sampling based methods (Zhou et al. 2020; Kang et al. 2020) over-sample tail classes to improve the representation for them. However, these two categories of methods are sub-optimal for untrimmed videos, as increasing the importance of tail classes may also bring background frame noise into training. Furthermore, it is revealed in existing long-tailed image recognition works (Cao et al. 2019; Cui et al. 2019; Kang et al. 2020) that both re-weighting and re-sampling based methods suffer from the performance drop on head classes when gaining the performance increase on tail classes. To alleviate this performance trade-off, mixture-of-experts models (Cai, Wang, and Hwang 2021; Wang et al. 2021b; Xiang, Ding, and Han 2020) are proposed for long-tailed image recognition, which learn a separate expert model for each group of samples.

With these observations in mind, we propose a Mixture-of-Experts framework with Internal Distillation (MEID) to tackle the long-tailed video recognition task with frame-level class imbalance. Specifically, the MEID framework is designed with two experts of different architectures, i.e., the first expert handling the original long-tailed distribution, and the second expert dealing with the re-balanced distribution from class-balanced sampling. In the first expert, a frame embedding module with a frame attention module to learn the importance of different frames is applied to aggregate temporal information for better video representation.

When designing the second expert, one practical issue arises in the video sample assignment stage. In the previous mixture-of-experts methods on long-tailed image recognition (Cai, Wang, and Hwang 2021; Wang et al. 2021b; Xiang, Ding, and Han 2020), different experts with the same architecture focus on handling image samples from different groups, i.e., head, medium and tail classes, to alleviate the class imbalance. However, in the case of multi-label videos, it is hard to assign one complete video to a single expert due to its multi-label property. Thus, we design the second expert with the frame-level assignment to alleviate the frame-level class imbalance. Specifically, the second expert contains a different architecture from the first expert, with a novel complementary frame selection module and a motion extraction module. The complementary frame selection module inherits the frame attention weights from the first expert, then selects frames with low attention weights and assigns them to the second expert, forcing the second expert to focus on the frames unsolved by the first expert. The motion extraction module extracts feature differences between adjacent frames for these selected frames to enhance their motion feature representation. Through the design of two experts with different architectures and the frame-level assignment, the model achieves the complementary effect between the two experts at the frame level.

Notably, in both experts, to improve the frame-level feature representation for tail classes and achieve the knowledge transfer from the head to tail classes, we modulate the frame-level feature with an auxiliary distillation loss to reduce the distribution distance between head and tail classes.

Despite that VideoLT (Zhang et al. 2021b) has multi-label merit, it only contains an average of 1.1 labels for each video. To highlight the multi-label challenge in long-tailed video recognition, we create two additional benchmarks based on Charades (Sigurdsson et al. 2016) and CharadesEgo (Sigurdsson et al. 2018) datasets with multi-label property, called CharadesLT and CharadesEgoLT.

The main contributions of this paper are as follows:

- We propose a mixture-of-experts framework for long-tailed video recognition considering the frame-level imbalance, and our two experts are designed with different architectures to focus on different frames within each video.
- To achieve frame-level assignment between the two experts, we design a complementary frame selection module in the second expert to select frames unsolved by the first expert based on attention weights.
- A novel head-to-tail distillation loss is proposed to improve the feature representation for tail classes.
- Extensive experiments are conducted on the VideoLT dataset, our proposed CharadesLT and CharadesEgoLT datasets to show the effectiveness of our proposed method with state-of-the-art performance.

Related Work

Long-Tailed Image Recognition

As a similar problem in the image domain, long-tailed image recognition has gained a lot of attention in recent years (Cai, Wang, and Hwang 2021; Wang et al. 2021b; Cui et al. 2021; Liu et al. 2021; Li, Wang, and Wu 2021; Ye, Zhan, and Chao 2021; Menon et al. 2021; Park et al. 2021; Zhong et al. 2021; Wang et al. 2021a; Zhang et al. 2021a). According to different paradigms, we briefly review them in the following sections:

Re-Weighting based Methods. This line of research (Cui et al. 2019; Cao et al. 2019; Tan et al. 2020) focuses on how to assign different losses to different samples from each class. Cui et al. (Cui et al. 2019) propose the class balanced (CB) loss with the theory of the effective number of samples to select samples for each class. Cao et al. (Cao et al. 2019) introduces a label-distribution-aware margin (LDAM) loss to enlarge margins for tail classes. Although loss re-weighting is easy to implement in practice as it does not require modification of the network, it suffers from performance drop on head classes.

Re-Sampling based Methods. Re-sampling based methods mostly rely on class-balanced sampling which typically oversamples samples from tail classes and discards samples from head classes. BBN (Zhou et al. 2020) explores the intrinsic mechanism of representation learning and classifier learning with sampling in long-tailed recognition. Kang et al. (Kang et al. 2020) work along the same direction and propose multiple sampling techniques to fine-tune the classifier. In this paper, we also use class-balanced sampling in the second expert to re-balance the distribution.

Feature Distribution Manipulation. Recently, various methods have been proposed to transfer feature representations of head classes to those of tail classes (Liu et al. 2020; Kim, Jeong, and Shin 2020) or directly augment tail classes (Li et al. 2021). GistNet (Liu et al. 2021) learns the

class geometry and transfers the geometry structure of head classes to tail classes. Liu et al. (Liu et al. 2020) propose the feature cloud, i.e., the intra-class angular distribution, which models the angular between each feature and the centre of each class, and then they reconstruct the feature cloud for tail classes. To modulate feature learning at the frame level, we design a distillation loss which transfers the prediction distribution of head classes to tail classes, instead of the angular distribution as in previous work (Liu et al. 2020).

Mixture-of-Experts Methods. Mixture-of-Experts (MoE) methods (Wang et al. 2021b; Xiang, Ding, and Han 2020) in long-tailed image recognition mainly distribute samples of head and tail classes to different experts and therefore mitigate the imbalance degree. LFME (Xiang, Ding, and Han 2020) merges all experts to the final model in a self-paced knowledge distillation manner. RIDE (Wang et al. 2021b) utilizes a diversity loss and a routing scheme to assign different samples to different experts. Different from these methods, our work does not explicitly assign samples to different experts considering the multi-label video setting, but focuses on the frame-level assignment with the selection of complementary frames.

Multi-Label Classification. Multi-label long-tailed image classification often suffers from the label co-occurrence problem. To solve this problem, Wu et al. (Wu et al. 2020) propose a distribution-balanced loss that re-balances the loss weights to alleviate the impact caused by label co-occurrence. Recently, Guo and Wang (Guo and Wang 2021) design a two-branch network and train each branch with uniform sampling and re-balanced sampling, respectively, aiming to learn a balanced implicit distribution. However, Guo and Wang (Guo and Wang 2021) choose to enforce the consistency of the two branches while our method increases the diversity of two different experts to achieve the complementary effect at the frame level.

Long-Tailed Video Recognition

Significant progress has been made in the video domain with various classification backbones and datasets (Carreira and Zisserman 2017; Lin, Gan, and Han 2019). Until recently, people start to investigate the long-tailed problem in video recognition and there exits only a few works currently. Zhu and Yang (Zhu and Yang 2020) simply extend their proposed image-based long-tailed model into the video domain without temporal modelling, and directly experiment on YouTube-8M (Abu-El-Haija et al. 2016) with pre-extracted features. Zhang et al. (Zhang et al. 2021b) propose FrameStack which performs frame-level sampling with the guidance of running average precision, without distinguishing the frame-level feature representation between head and tail classes. MEDC (Hu, Gao, and Xu 2022) focuses on calibrating the intra-class distribution to adapt to different inter-class distributions in the test phase without considering the frame-level imbalance in a multi-label setting. We consider the frame-level imbalance with frame-level feature learning in our proposed model.

Long-Tailed Video Datasets. Among the video datasets (Abu-El-Haija et al. 2016; Zhang et al. 2021b; Sigurdsson et al. 2016) , VideoLT (Zhang et al. 2021b) is specifically constructed for the long-tailed video recognition task. VideoLT exhibits multi-label property but is still sparse in terms of the number of labels per video with an average number of 1.1. To highlight the multi-label challenge in long-tailed video recognition, we propose two new benchmarks based on the videos from Charades (Sigurdsson et al. 2016) and CharadesEgo (Sigurdsson et al. 2018) datasets, named CharadesLT and CharadesEgoLT. Charades and CharadesEgo datasets contain multi-label videos with dense action annotations. Charades is mostly daily indoor activities with annotations for action recognition and temporal action detection. CharadesEgo dataset shares the same label space with Charades but the videos are recorded from both third-person and first-person views.

Method

In this section, we introduce the details of our proposed Mixture-of-Experts framework with Internal Distillation (MEID) for long-tailed video recognition. The long-tailed video recognition is formulated as follows. The input video data is denoted as $\mathbf{V} \in \mathbb{R}^{B\times N}$, where B is batch size and N indicates the number of frames for each video. The corresponding label set is denoted as $\mathbf{Y} \in \mathbb{R}^{B\times C}$ and C represents the number of classes.

The overall framework of MEID is illustrated in Fig. 2. The visual features are firstly extracted from the pre-trained visual encoder and then highlighted with predicted attention weights. On top of the visual feature embedding, to further improve the feature representation for tail classes, we propose a novel distillation loss to reduce the distribution distance between head and tail classes. To tackle the multi-label challenge for long-tailed video recognition, we design a frame-level mixture-of-experts model with each expert focusing on different frames from each video, which is realized by a complementary frame selection module.

Attention Based Feature Extraction

Pre-trained Visual Encoder. The input videos $\mathbf{V}$ first go through the pre-trained visual encoder to get visual features $\mathbf{X}$, as shown below:

$$\mathbf{X} = VisualEncoder(\mathbf{V}) \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{B\times N\times D}$ and D indicates the visual feature dimension, which is typically the output feature from the last convolution layer. Following the previous long-tailed video recognition work VideoLT (Zhang et al. 2021b), we also evaluate various pre-trained classification backbones, including ResNet50, ResNet101 (He et al. 2016) and TSM-ResNet50 (Lin, Gan, and Han 2019).

Frame Embedding. The extracted pre-trained visual features $\mathbf{X}$ are further embedded to learn task-related features, as shown below:

$$\mathbf{T} = FrameEmbedding(\mathbf{X}) \tag{2}$$

where the output from the frame embedding module is indicated as $\mathbf{T} \in \mathbb{R}^{B\times N\times D}$. Specifically, the frame embedding module is designed as a sequence of operators consisting of normalization, linear layer and relu activation.

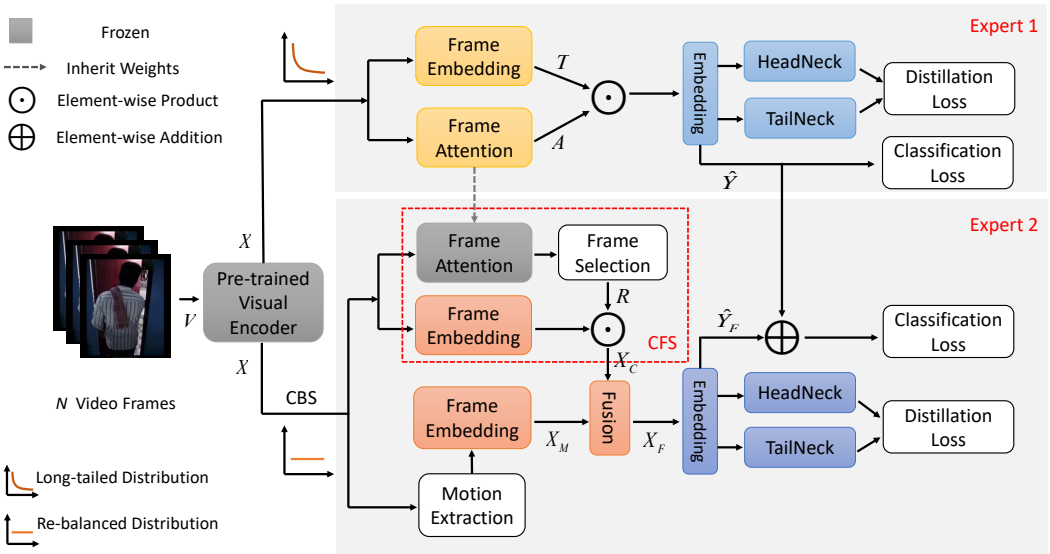

Figure 2: Overview of our MEID framework with two experts focusing on different frames of the input video for long-tailed video recognition. Given N video frames, the frame features are extracted from the pre-trained visual encoder. The first expert handles the original long-tailed distribution with a frame embedding module and a frame attention module, optimized by the classification loss and our proposed distillation loss applied between two embedding necks of head and tail classes to enhance the feature representation of tail classes. The second expert deals with the re-balanced distribution after class-balanced sampling (CBS), equipped with a complementary frame selection (CFS) module to focus on frames unsolved by the first expert, and a motion extraction module to extract additional motion features. The second expert inherits the frame attention module from the first expert (indicated by the grey dashed arrow), and also integrates the classification embedding from the first expert into its classification loss during training as well as the final video prediction during inference.

Frame Attention. To indicate the importance of each frame, we learn the frame attention weights $\mathbf{A}$ as shown below, with two linear layers connected from pre-trained visual features $\mathbf{X}$.

$$\mathbf{A} = FrameAttention(\mathbf{X}) \tag{3}$$

where the frame attention weights $\mathbf{A} \in \mathbb{R}^{B\times N}$. Notably, the frame attention weights $\mathbf{A}$ also serve one important role of frame-level assignment in the two experts.

Classification Embedding. The frame attention weights $\mathbf{A}$ will apply on the frame embedding $\mathbf{T}$ to highlight selected frames in the final video representation, before classification embedding $\hat{\mathbf{Y}}$, as shown below:

$$\hat{\mathbf{Y}} = Linear(\mathbf{T} \odot \mathbf{A}) \tag{4}$$

Distillation Loss

To improve the frame-level feature representation of tail classes and achieve the knowledge transfer from head classes to tail classes, we propose a novel distillation loss based on the statistics of head classes and tail classes. Firstly, the head classes part $\hat{\mathbf{Y}}_{head}$ and the tail classes part $\hat{\mathbf{Y}}_{tail}$ are split from the whole label space $\hat{\mathbf{Y}}$ with masking. Then, a distance loss is applied between the statistics of the head classes $\hat{\mathbf{Y}}_{head}$ and the tail classes $\hat{\mathbf{Y}}_{tail}$ to realize head-to-tail knowledge transfer. Considering that the head and tail classes are two disjoint sets, we use variance value instead of mean value as the mean value is class-dependent.

Practically, directly applying the distillation loss between the variance values of the head classes $\hat{\mathbf{Y}}_{head}$ and the tail classes $\hat{\mathbf{Y}}_{tail}$ might disturb the normal video classification resulting in the inability to optimize. To alleviate this problem, we add one more embedding (indicated as **neck**) for head classes $\hat{\mathbf{Y}}_{head}$ and tail classes $\hat{\mathbf{Y}}_{tail}$ before distillation, so that the proposed distillation loss and the original video classification loss are optimized in different feature spaces with weak coupling.

The necks M_{head} and M_{tail}, and the distillation loss L_d based on distance calculation are described as follows:

$$\mathrm{M_{head}}(\hat{\mathbf{Y}}) = W_{head} * (\hat{\mathbf{Y}}_{head}) + b_{head} \tag{5}$$

$$\mathrm{M_{tail}}(\hat{\mathbf{Y}}) = W_{tail} * (\hat{\mathbf{Y}}_{tail}) + b_{tail} \tag{6}$$

$$\mathrm{L_d} = \|(Var(\mathrm{M_{head}}(\hat{\mathbf{Y}})) - Var(\mathrm{M_{tail}}(\hat{\mathbf{Y}})))\|_2 \tag{7}$$

where W and b represent the embedding weights and bias in the corresponding neck for the head and tail.

Mixture-of-Experts Model

As shown in Fig. 2, our MEID framework consists of two experts with different architectures, i.e., the first expert handling the original long-tailed distribution, and the second expert dealing with the re-balanced distribution from class-balanced sampling. Through the frame-level assignment, we maximize the diversity between the two experts and realize the complementary effect between them.

Notably, the basic feature extraction backbone consists of the frame embedding with the frame attention indicating the frame importance and is applied in both experts to aggregate temporal information for better video representation. When designing the second expert, video samples are applied with **class-balanced sampling** to get one re-balanced distribution, compared to the original long-tailed distribution in the first expert. Besides the re-balanced data distribution, the second expert also contains two specially designed modules, i.e., the Complementary Frame Selection and the Motion Extraction, to select unsolved frames by the first expert and enhance their motion feature representation.

Complementary Frame Selection The Mixture-of-Experts (MoE) mechanism has been proven effective in long-tailed image recognition (Xiang, Ding, and Han 2020; Wang et al. 2021b), where different experts handle different subsets of long-tailed images and the imbalance ratio is decreased in each expert. However, this procedure can not directly adapt to the long-tailed video task as it is hard to directly assign one complete video to only one expert considering the multi-label fact.

In our work, to adapt the mixture-of-experts model to the multi-label setting in videos, we resort to the frame-level assignment and make the second expert focus on frames unsolved by the first expert. Specifically, we inherit and reuse the frame attention weights from the first expert, and we calculate the reciprocal of the frame attention weights to suppress the easy frames with high attention weights and force the second expert to learn the rest hard frames. Through these two experts, the model can learn video representation in a complementary way. The process of generating the complementary frame attention weights $\mathbf{R}$ in the second expert conditioned on the first expert's frame attention weights $\mathbf{A}$ is as follows:

$$\mathbf{R} = \mathbf{1}/(\mathbf{A} + \beta) \tag{8}$$

where $\beta \in (0, 1)$ indicates the hyperparameter to avoid the extreme value of R. The complementary frame attention weights $\mathbf{R}$ is applied on the frame embedding of pre-extracted visual features $\mathbf{X}$ to highlight complementary frames in the final video representation $\mathbf{X_C}$, similar to that in the first expert, as shown below:

$$\mathbf{X_C} = FrameEmbedding(\mathbf{X}) \odot \mathbf{R} \tag{9}$$

Motion Extraction Some video segments are action-oriented with action dominating (e.g., 'tango'), while other video segments are object-oriented with object dominating (e.g., 'bear'). Motion modelling is typically harder than object appearance recognition due to the temporal evolution of motion feature. Considering that our second expert tackles the hard frames unsolved by the first expert which might contain more action-oriented video segments, we incorporate the motion feature into the second expert.

Specifically, we get the motion feature by calculating the feature differences between adjacent frames. To maintain the static object information in the final motion feature representation $\mathbf{X_M}$, we also compute the mean value of all frames' visual features $\mathbf{X} = \{x_n, n \in 1, 2, ..., N\}$, and concatenate it with the frame feature differences, formulated as follows:

$$\mathbf{X_M} = Concat(x_2 - x_1, ..., x_n - x_{n-1}, (\sum_{n=1}^{N} x_n)/N) \tag{10}$$

Next, the frame embedding of motion feature $\mathbf{X_M}$ is fused with the complementary video feature representation $\mathbf{X_C}$ using a simple concatenation and linear projection module:

$$\mathbf{X_F} = Linear(Concat(X_C, FrameEmbedding(X_M))) \tag{11}$$

On the fused video feature $\mathbf{X_F}$, the classification **embedding** layer is applied for the second expert:

$$\hat{\mathbf{Y}}_\mathbf{F} = Linear(\mathbf{X_F}) \tag{12}$$

Optimization

In the training phase, we use a two-stage training strategy. Firstly, we train the first expert combining a classification loss L_{cls} and our proposed distillation loss L_d, as shown below:

$$\text{Loss1} = L_{cls}(\hat{\mathbf{Y}}) + \lambda_1 L_d(\mathrm{M}_{\mathrm{head}}(\hat{\mathbf{Y}}), \mathrm{M}_{\mathrm{tail}}(\hat{\mathbf{Y}})) \tag{13}$$

Secondly, we freeze the first expert and inherit only the frame attention module in Eq. 3 to train the second expert. Notably, we also freeze the frame attention module in the second expert to guarantee the consistency of the frame attention weights $\mathbf{A}$ in the two experts. To retain the information from the first expert and achieve the complementary effect between the two experts, the classification loss L_{cls} in the second expert accepts the addition of the two classification embeddings from the two experts, i.e., $\hat{\mathbf{Y}}$ and $\hat{\mathbf{Y}}_\mathbf{F}$. The overall loss in the second expert is shown as follows:

$$\text{Loss2} = L_{cls}(\hat{\mathbf{Y}} + \hat{\mathbf{Y}}_\mathbf{F}) + \lambda_2 L_d(\mathrm{M}_{\mathrm{head}}(\hat{\mathbf{Y}}_\mathbf{F}), \mathrm{M}_{\mathrm{tail}}(\hat{\mathbf{Y}}_\mathbf{F})) \tag{14}$$

In the inference phase, based on the addition of the two classification embeddings in the two experts, i.e., $\hat{\mathbf{Y}}$ and $\hat{\mathbf{Y}}_\mathbf{F}$, the final video labels are predicted.

Experiments

Datasets

VideoLT. VideoLT (Zhang et al. 2021b) is the only existing long-tailed video recognition benchmark which covers various daily activities. In this dataset, ResNet-50 (He et al. 2016) and ResNet-101 pre-trained on ImageNet, and TSM-ResNet-50 (Lin, Gan, and Han 2019) backbone pre-trained on Kinetics (Carreira and Zisserman 2017) are used as pre-trained visual encoders to extract features from raw videos. We conduct experiments using the three categories of pre-extracted features. Each video contains 150 frames and each frame contains 2048 dimensional features.

CharadesLT and CharadesEgoLT. Though VideoLT (Zhang et al. 2021b) has the multi-label property, the

Datasets	#Classes	#Head	#Medium	#Tail	#Training	#Validation	#Test	#Label/#Video	#IR
VideoLT (Zhang et al. 2021b)	1,004	47	617	340	179,334	25619	51,239	1.1	43.5
CharadesLT (ours)	157	41	81	35	2,213	469	1,012	**6.9**	**61.8**
CharadesEgoLT (ours)	157	43	89	25	1,553	764	748	**9.6**	**53.8**

Table 1: Data statistics on three datasets. #Label/#Video represents the average number of labels in each video. #IR represents the imbalance ratio which is the quotient of the maximum and minimum numbers of videos per class.

M	V	A	Overall mAP	Head (500,+∞)	Medium (100,500]	Tail (0,100]
LDAM	ResNet50	Nonlinear	0.502	0.680	0.557	0.378
EQL			0.502	0.679	0.557	0.378
CBS			0.491	0.649	0.545	0.371
CBL			0.495	0.653	0.546	0.381
Mixup			0.484	0.649	0.535	0.368
FS			0.516	0.683	0.569	0.397
MEDC			0.567	0.720	0.607	0.436
Ours			**0.619**	**0.750**	**0.661**	**0.524**
LDAM	ResNet101	Nonlinear	0.518	0.687	0.572	0.397
EQL			0.518	0.690	0.571	0.398
CBS			0.507	0.660	0.559	0.390
CBL			0.511	0.665	0.561	0.398
Mixup			0.500	0.665	0.550	0.386
FS			0.535	0.697	0.587	0.419
MEDC			0.603	0.737	0.657	0.499
FS		NV	0.670	0.780	0.707	0.590
Ours			**0.716**	**0.811**	**0.745**	**0.652**

Table 2: Experimental results on the VideoLT dataset with ResNet-50 and ResNet-101 pre-trained visual encoders. The existing model results are from VideoLT (Zhang et al. 2021b) and MEDC (Hu, Gao, and Xu 2022). 'M': Methods; 'V': Visual Encoder; 'A': Aggregation; 'NV': NetVLAD; 'CBL': CB Loss; 'FS': FrameStack.

	Methods	Overall mAP	Head (100,+∞)	Medium (20,100]	Tail (0,20]
CharadesLT	LDAM	0.114	0.191	0.105	0.046
	EQL	0.121	0.203	0.113	0.042
	CBS	0.129	0.208	0.124	0.047
	CB Loss	0.112	0.162	0.113	0.052
	Mixup	0.119	0.198	0.113	0.043
	FrameStack	0.113	0.194	0.103	0.042
	Ours	**0.151**	**0.217**	**0.151**	**0.076**
CharadesEgoLT	LDAM	0.154	0.230	0.137	0.082
	EQL	0.158	0.250	0.136	0.076
	CBS	0.163	0.255	0.142	0.079
	CB Loss	0.129	0.190	0.115	0.077
	Mixup	0.155	0.243	0.133	0.081
	FrameStack	0.152	0.237	0.131	0.080
	Ours	**0.197**	**0.280**	**0.179**	**0.118**

Table 3: Experimental results on CharadesLT and CharadesEgoLT datasets. NeXtVLAD is used for feature aggregation, and ResNet-101 is used as the pre-trained visual encoder.

average number of labels per video is 1.1, still not showing the real challenges in practical applications. We propose two new benchmarks, CharadesLT and CharadesEgoLT, based on the videos from Charades dataset (Sigurdsson et al. 2016) and CharadesEgo (Sigurdsson et al. 2018), which contains more labels for each video. ImageNet pre-trained ResNet-101 model is used as our pre-trained visual encoder to extract features for the uniformly sampled 200 frames from each video, i.e., the 2048 dimensional features from the penultimate layer.

The data statistics of three datasets are shown in Table 1. Compared to VideoLT, our proposed datasets highlight the multi-label challenge with more labels in each video and a higher imbalance ratio.

Experimental Settings

Implementation Details. The Adam optimizer with an initial learning rate of 1e-3 is used to train the first stage. In the second stage, the initial learning rate of 1e-4 is used. Each stage is trained for 100 epochs, and the learning rate decayed at the 30th and 60th epochs by 0.1. We adopt focal loss (Lin et al. 2020) as our classification loss following the previous work (Zhang et al. 2021b). We set the loss weights λ_1 in Eq. 13 and λ_2 in Eq. 14 to 0.1. Last, all the experiments are conducted on the PyTorch framework with NVIDIA A5000 GPUs.

Evaluation Metrics. We evaluate our method via mean average precision (mAP) on all classes. Additionally, we report per-group mAP on head classes, medium classes and tail classes to better reflect the performance breakdown of our method.

Comparison with State-of-the-art Methods

Baselines. The VideoLT work (Zhang et al. 2021b) proposes the FrameStack model and establishes various baseline results. We list these results and the recent MEDC (Hu, Gao, and Xu 2022) result in Table 2, and compare our method with them. Models with various pre-trained visual encoders, and video-level feature aggregation methods like Nonlinear aggregation (Zhang et al. 2021b), NetVLAD (Arandjelovic et al. 2018) and NextVLAD (Lin, Xiao, and Fan 2018), are explored to further compare the performance.

Results Table 2 shows that our method outperforms all previous methods by a large margin on the VideoLT (Zhang et al. 2021b) dataset. Under ResNet-50 (He et al. 2016) backbone, our model outperforms the previous best result (MEDC (Hu, Gao, and Xu 2022)) by 5.2% mAP. The FrameStack model (abbreviated as FS in Table 2) with

FA	DL	Overall mAP	Head (500,+∞)	Medium (100,500]	Tail (0,100]
		0.564	0.710	0.622	0.439
✓		0.580	0.715	0.629	0.472
✓	✓	0.588	0.720	0.636	0.483

Table 4: Ablation studies for Frame Attention (FA) and Distillation Loss (DL). Experiments are conducted on the first expert.

CFS	ME	CBS	Overall mAP	Head (500,+∞)	Medium (100,500]	Tail (0,100]
			0.588	0.720	0.636	0.483
✓			0.604	0.734	0.651	0.501
✓	✓		0.615	0.757	0.661	0.513
✓	✓	✓	0.619	0.750	0.661	0.524

Table 5: Ablation studies for the Complementary Frame Selection (CFS), Motion Extraction (ME) and Class-Balanced Sampling (CBS) designs in the second expert.

NetVLAD (Arandjelovic et al. 2018) feature aggregation outperforms that with Nonlinear aggregation (Zhang et al. 2021b), and becomes the best performing model using ResNet-101 backbone. However, under the same ResNet-101 backbone and NetVLAD aggregation, our method still outperforms FrameStack by 4.6% mAP. Notably, the performance of our method on tail classes improves more than that on the head and medium classes, demonstrating the superiority of our method on the long-tailed video recognition task.

The baseline results and our proposed model results on CharadesLT (Sigurdsson et al. 2016) and CharadesEgoLT (Sigurdsson et al. 2018) datasets are shown in Table 3. While our model outperforms all the existing model results, the values are still relatively low, showing the challenges of our proposed two datasets. In addition to the multi-label challenge with a higher imbalance ratio, the classes in Charades and CharadesEgo datasets are mostly action-oriented with action dominating, which brings an extra challenge in feature encoding.

Ablation Studies

The ablation studies are conducted on the VideoLT dataset with ResNet-50 as the pre-trained visual encoder.

Frame Attention and Distillation Loss. In Table 4, applying the frame attention to indicate the frame importance or the distillation loss to transfer the knowledge from head to tail classes can both bring performance improvement in all evaluation metrics. And the combination usage can further improve the results. Notably, the improvements in the tail class are more significant than that in head classes, indicating the benefits of these two modules for the long-tailed problem.

Complementary Frame Selection (CFS), Motion Extraction (ME) and Class-Balanced Sampling (CBS). Table 5 shows that each of the three components (CFS, ME and CBS) can gradually bring result improvement. Notably, our designed CFS module for frame-level assignment between the two experts achieves a complementary effect and improves the tail class performance significantly.

	Overall mAP	Head (500,+∞)	Medium (100,500]	Tail (0,100]
E1	0.588	0.720	0.636	0.483
E1+E1	0.592	0.723	0.650	0.467
E1+E2	0.619	0.750	0.661	0.524

Table 6: Ablation studies for two-expert architecture. 'E1': Expert One; 'E2': Expert Two.

β	Overall mAP	Head (500,+∞)	Medium (100,500]	Tail (0,100]
0.3	0.611	0.746	0.651	0.521
0.5	0.619	0.750	0.661	0.524
0.7	0.612	0.747	0.655	0.514

Table 7: Ablation studies for hyperparameter β in CFS.

Two-Expert Architecture. In Table 6, the full two-expert model 'E1+E2' improves significantly over the first expert 'E1'. To exclude the influence of extra model parameters, we construct another baseline 'E1+E1' with two same 'E1' in both branches. Compared to 'E1+E2', the performance of 'E1+E1' drops significantly in all metrics reflecting the superiority of our designed second expert. Notably, tail class performance in 'E1+E1' even drop by 1.6% mAP compared to 'E1', indicating that adding one more expert without our designed modules may still focus more on easy frames typically from head classes and ignore the hard frames normally from tail classes.

Hyperparameter β. The effect of hyperparameter β in the CFS module to avoid the extreme value of complementary frame attention weights is shown in Table 7 with 0.5 as the best choice.

Conclusion

In this paper, we propose a Mixture-of-Experts framework with Internal Distillation (MEID) for the long-tailed video recognition task. The MEID framework contains two experts with different architectures to focus on different frames within each video. Frames unsolved by the first expert are selected through our designed complementary frame selection module in the second expert and further enhanced with motion features. A novel head-to-tail distillation loss is designed to improve the frame-level feature representation for tail classes. We also propose two new benchmarks, CharadesLT and CharadesEgoLT, which are more imbalanced and contain more multi-label videos, to demonstrate real-world challenges and foster future research in this direction. State-of-the-art performance is achieved on VideoLT and our proposed benchmarks. In the future, the long-tailed problem in other video tasks, e.g., action localization, could also be explored.

References

Abu-El-Haija, S.; Kothari, N.; Lee, J.; Natsev, P.; Toderici, G.; Varadarajan, B.; and Vijayanarasimhan, S. 2016. YouTube-8M: A Large-Scale Video Classification Benchmark. *CoRR*, abs/1609.08675.

Arandjelovic, R.; Gronát, P.; Torii, A.; Pajdla, T.; and Sivic, J. 2018. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6): 1437–1451.

Cai, J.; Wang, Y.; and Hwang, J.-N. 2021. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 112–121.

Cao, K.; Wei, C.; Gaidon, A.; Aréchiga, N.; and Ma, T. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 1565–1576.

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 4724–4733. IEEE Computer Society.

Cui, J.; Zhong, Z.; Liu, S.; Yu, B.; and Jia, J. 2021. Parametric contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 715–724.

Cui, Y.; Jia, M.; Lin, T.; Song, Y.; and Belongie, S. J. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 9268–9277. Computer Vision Foundation / IEEE.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. SlowFast Networks for Video Recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 6201–6210. IEEE.

Guo, H.; and Wang, S. 2021. Long-Tailed Multi-Label Visual Recognition by Collaborative Training on Uniform and Re-Balanced Samplings. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 15089–15098. Computer Vision Foundation / IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778. IEEE Computer Society.

Hu, Y.; Gao, J.; and Xu, C. 2022. Learning Muti-expert Distribution Calibration for Long-tailed Video Classification. *CoRR*, abs/2205.10788.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2020. Decoupling Representation and Classifier for Long-Tailed Recognition. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Kim, J.; Jeong, J.; and Shin, J. 2020. M2m: Imbalanced Classification via Major-to-Minor Translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 13893–13902. Computer Vision Foundation / IEEE.

Li, S.; Gong, K.; Liu, C. H.; Wang, Y.; Qiao, F.; and Cheng, X. 2021. MetaSAug: Meta Semantic Augmentation for Long-Tailed Visual Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 5212–5221. Computer Vision Foundation / IEEE.

Li, T.; Wang, L.; and Wu, G. 2021. Self supervision to distillation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 630–639.

Lin, J.; Gan, C.; and Han, S. 2019. TSM: Temporal Shift Module for Efficient Video Understanding. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 7082–7092. IEEE.

Lin, R.; Xiao, J.; and Fan, J. 2018. NeXtVLAD: An Efficient Neural Network to Aggregate Frame-Level Features for Large-Scale Video Classification. In Leal-Taixé, L.; and Roth, S., eds., *Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, volume 11132 of *Lecture Notes in Computer Science*, 206–218. Springer.

Lin, T.; Goyal, P.; Girshick, R. B.; He, K.; and Dollár, P. 2020. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2): 318–327.

Liu, B.; Li, H.; Kang, H.; Hua, G.; and Vasconcelos, N. 2021. GistNet: a Geometric Structure Transfer Network for Long-Tailed Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8209–8218.

Liu, J.; Sun, Y.; Han, C.; Dou, Z.; and Li, W. 2020. Deep Representation Learning on Long-Tailed Data: A Learnable Embedding Augmentation Perspective. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2967–2976. Computer Vision Foundation / IEEE.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2021. Long-tail learning via logit adjustment. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Park, S.; Lim, J.; Jeon, Y.; and Choi, J. Y. 2021. Influence-balanced loss for imbalanced visual classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 735–744.

Sigurdsson, G. A.; Gupta, A.; Schmid, C.; Farhadi, A.; and Alahari, K. 2018. Actor and Observer: Joint Modeling of First and Third-Person Videos. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 7396–7404. Computer Vision Foundation / IEEE Computer Society.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, volume 9905 of *Lecture Notes in Computer Science*, 510–526. Springer.

Tan, J.; Wang, C.; Li, B.; Li, Q.; Ouyang, W.; Yin, C.; and Yan, J. 2020. Equalization Loss for Long-Tailed Object Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 11659–11668. Computer Vision Foundation / IEEE.

Wang, P.; Han, K.; Wei, X.; Zhang, L.; and Wang, L. 2021a. Contrastive Learning Based Hybrid Networks for Long-Tailed Image Classification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 943–952. Computer Vision Foundation / IEEE.

Wang, X.; Lian, L.; Miao, Z.; Liu, Z.; and Yu, S. X. 2021b. Long-tailed Recognition by Routing Diverse Distribution-Aware Experts. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Wu, T.; Huang, Q.; Liu, Z.; Wang, Y.; and Lin, D. 2020. Distribution-Balanced Loss for Multi-label Classification in Long-Tailed Datasets. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, volume 12349 of *Lecture Notes in Computer Science*, 162–178. Springer.

Xiang, L.; Ding, G.; and Han, J. 2020. Learning From Multiple Experts: Self-paced Knowledge Distillation for Long-Tailed Classification. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V*, volume 12350 of *Lecture Notes in Computer Science*, 247–263. Springer.

Ye, H.-J.; Zhan, D.-C.; and Chao, W.-L. 2021. Procrustean training for imbalanced deep learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 92–102.

Zhang, S.; Li, Z.; Yan, S.; He, X.; and Sun, J. 2021a. Distribution Alignment: A Unified Framework for Long-Tail Visual Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 2361–2370. Computer Vision Foundation / IEEE.

Zhang, X.; Wu, Z.; Weng, Z.; Fu, H.; Chen, J.; Jiang, Y.-G.; and Davis, L. S. 2021b. VideoLT: Large-scale Long-tailed Video Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7960–7969.

Zhong, Z.; Cui, J.; Liu, S.; and Jia, J. 2021. Improving Calibration for Long-Tailed Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 16489–16498. Computer Vision Foundation / IEEE.

Zhou, B.; Cui, Q.; Wei, X.; and Chen, Z. 2020. BBN: Bilateral-Branch Network With Cumulative Learning for Long-Tailed Visual Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 9716–9725. Computer Vision Foundation / IEEE.

Zhu, L.; and Yang, Y. 2020. Inflated Episodic Memory With Region Self-Attention for Long-Tailed Visual Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 4343–4352. Computer Vision Foundation / IEEE.