Generalized Many-Way Few-Shot Video Classification

Yongqin Xian[1], Bruno Korbar[2], Matthijs Douze[2], Bernt Schiele[1],
Zeynep Akata[1,3], and Lorenzo Torresani[2]

[1] Max Planck Institute for Informatics
[2] Facebook AI
[3] University of Tübingen

Abstract. Few-shot learning methods operate in low data regimes. The aim is to learn with few training examples per class. Although significant progress has been made in few-shot image classification, few-shot video recognition is relatively unexplored and methods based on 2D CNNs are unable to learn temporal information. In this work we thus develop a simple 3D CNN baseline, surpassing existing methods by a large margin. To circumvent the need of labeled examples, we propose to leverage weakly-labeled videos from a large dataset using tag retrieval followed by selecting the best clips with visual similarities, yielding further improvement. Our results saturate current 5-way benchmarks for few-shot video classification and therefore we propose a new challenging benchmark involving more classes and a mixture of classes with varying supervision.

1 Introduction

In the video domain annotating data is very time-consuming due to the additional time dimension. The lack of labeled training data is more prominent for some fine-grained action classes at the "tail" of the skewed long-tail distribution (see Figure 1), e.g., "arabesque in ballet". It is thus important to learn to classify videos in the limited labeled training data regime. Visual recognition methods that operate in the few-shot learning setting aim to generalize a classifier trained on *base classes* with enough training data to *novel classes* with only a few labeled training examples. While considerable attention has been devoted to this scenario in the image domain [41,28,29,4], few-shot video classification is relatively unexplored.

Existing few-shot video classification approaches [48,2] are mostly based on frame-level features extracted from a 2D CNN, which essentially ignores the important temporal information. Although additional temporal modules have been added at the top of a pre-trained 2D CNN, necessary temporal cues may be lost when temporal information is learned on top of static image features. We argue that under-representing temporal cues may negatively impact the robustness of the classifier. In fact, in the few-shot scenario it may be risky for the model to rely exclusively on appearance and context cues extrapolated from the few available examples. In order to make temporal information available we propose to represent the videos by means of a 3D CNN.

While obtaining labeled videos for target classes is time-consuming and challenging, there are many videos tagged by users available on the internet. For example, there are 400,000 tag-labeled videos in the YFCC100M [36] dataset. Our second goal is thus

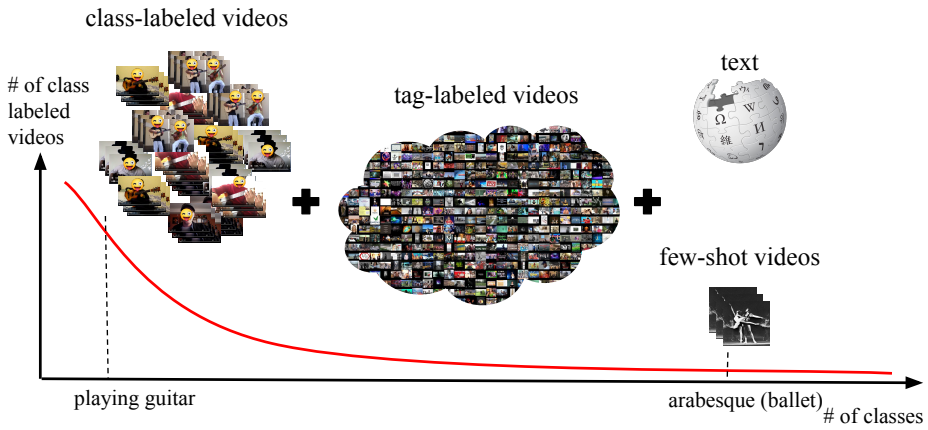

Fig. 1. Our 3D CNN approach combines a few class-labeled videos (time-consuming to obtain) and tag-labeled videos. It saturates existing benchmarks, so we move to a more challenging generalized many-way few-shot video classification task.

to leverage such tag-labeled videos (Figure 1) to alleviate the lack of labeled training data.

Existing experimental settings for few-shot video classification [48,2] are limited. Predicting a label among just 5 novel classes in each testing episode is in fact relatively easy. Moreover, restricting the label space to only novel classes at test time, and ignoring the base classes is unrealistic. In real-world applications test videos are expected to belong to any class.

In this work, our goal is to push the progress of few-shot video classification in three ways: 1) To learn the temporal information, we revisit spatiotemporal CNNs in the few-shot video classification regime. We develop a 3D CNN baseline that maintains significant temporal information within short clips; 2) We propose to retrieve relevant videos annotated with tags from a large video dataset (YFCC100M) to circumvent the need for labeled videos of novel classes; 3) We extend current few-shot video classification evaluation settings by introducing two challenges. In our *generalized few-shot video classification* task, the label space has no restriction in terms of classes. In many-way few-shot video classification with, the number of classes goes well beyond five, and towards all available classes. Our extensive experimental results demonstrate that on existing settings spatiotemporal CNNs outperform the state-of-the-art by a large margin, and on our proposed settings weakly-labeled videos retrieved using tags successfully tackles both of our new few-shot video classification tasks.

2 Related work

Low-shot learning setup. The low-shot image classification [26,29,15] setting uses a large-scale fully labeled dataset for pre-training a DNN on the base classes, and a low-shot dataset with a small number of examples from a disjoint set of novel classes. The

terminology "k-shot n-way classification" means that in the low-shot dataset there are n distinct classes and k examples per class for training. Evaluating with few examples (k small) is bound to be noisy. Therefore, the k training examples are often sampled several times and accuracy results are averaged [15,6]. Many authors focus on cases where the number of classes n is small as well, which amplifies the measurement noise. For that case [29] introduces the notion of "episodes". An episode is one sampling of n classes and k examples per class, and the accuracy measure is averaged over episodes.

It is feasible to use distinct datasets for pre-training and low-shot evaluation. However, to avoid dataset bias [37] it is easier to split a large supervised dataset into disjoint sets of "base" and "novel" classes. The evaluation is often performed only on novel classes, except [15,45,32] who evaluate on the combination of base+novel classes.

Recently, a low-shot video classification setup has been proposed [48,7]. They use the same type of decomposition of the dataset as [29], with learning episodes and random sampling of low-shot classes. In this work, we follow and extend the evaluation protocol of [48].

Tackling low-shot learning. The simplest low-shot learning approach is to extract embeddings from the images using the pre-trained DNN and train a linear classifier [1] or logistic regression [15] on these embeddings using the k available training examples. Another approach is to cast low-shot learning as a nearest-neighbor classifier [44]. The "imprinting" approach [28], consists in building a linear classifier from the embeddings of training examples, then fine-tune it. Note that this is close to a nearest-neighbor classifier, since it is equivalent to doing class-mean similarity search with a cosine distance. As a complementary approach, [19] has looked into exploiting noisy labels to aid classification. By leveraging tags of 100M images from the YFCC100M dataset [36], they show improvements over Imagenet-pretraining. In this work, we use videos from YFCC100M retrieved by tags to augment and improve training of our few-shot classifier.

In a meta-learning setup, the low-shot classifier is assumed to have hyper-parameters or parameters that must be adjusted before training. Thus, there is a preliminary meta-learning step that consists in training those parameters on simulated episodes sampled from the base classes. Both Matching networks [41] and Prototypical Networks[34] employ metric learning to "meta-learn" deep neural features and adopt a nearest neighbor classifier. [29] meta-learns the optimization algorithm via an LSTM that maps the low-shot training examples into a classifier. Feature hallucination [15] meta-learns how to generate additional training data for novel classes, directly in the feature space. In MAML [11], the embedding classifier is meta-learned to adapt quickly and without overfitting to fine-tuning. Ren et al. [30] introduce a semi-supervised meta-learning approach that includes unlabeled examples in each training episode. While that method holds out a subset from the same target dataset as the unlabeled images, our retrieval-enhanced approach leverages weakly-labeled videos from another heterogeneous dataset which may have domain shift issues and a huge amount of distracting videos.

Recent works [4,44] suggest that state-of-the-art performance can be obtained by methods that do not need meta learning. In particular, [4] show that meta-learning methods are less useful when the image descriptors are expressive enough, which is the case

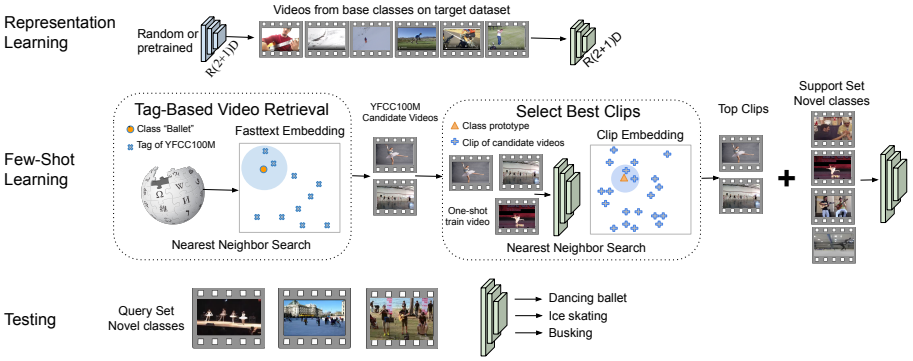

Fig. 2. Our approach comprises three steps: representation learning, few-shot learning and testing. In representation learning, we train a R(2+1)D CNN using the base classes of our target dataset starting from a random initialization or from a Sports1M-pretrained model. In few-shot learning, given few-shot support videos from novel classes, we first retrieve a list of candidate videos for each class from YFCC100M [36] using their tags, followed by selecting the best matching short clips from the retrieved videos using visual features. Those clips serve as additional training examples to learn classifiers that generalize to novel classes at test time.

when they are from high-capacity networks trained on large datasets. Therefore, we focus on techniques that do not require a meta-learning stage.

Deep descriptors for videos. Moving from hand-designed descriptors [5,24,31,42] to learned deep network based descriptors [9,10,21,33,43,38] has been enabled by labeled large-scale datasets [22,21], and parallel computing hardware. Deep descriptors are sometimes based on 2D-CNN models operating on a frame-by-frame basis with temporal aggregation [13,47]. More commonly they are 3D-CNN models that operate on short sequences of images that we refer to as video-clips [38,40]. Recently, ever-more-powerful descriptors have been developed leveraging two-stream architectures using additional modalities [10,33], factorized 3D convolutions [40,39], or multi-scale approaches [8]. While most of these descriptors are trained in a fully supervised way, advances in learning deep descriptors in either weakly supervised [46,12,25] or self supervised fashion have been explored as well [23,27].

3 Learning spatiotemporal features of videos

In the few-shot learning setting [48], classes are split into two disjoint label sets, i.e., base classes (denoted as $\mathbf{C}_b$) that have a large number of training examples, and novel classes (denoted as $\mathbf{C}_n$) that have only a small set of training examples. Let $\mathbf{X}_b$ denote the training videos with labels from the base classes and $\mathbf{X}_n$ be the training videos with labels from the novel classes ($|\mathbf{X}_b| \gg |\mathbf{X}_n|$). Given the training data $\mathbf{X}_b$ and $\mathbf{X}_n$, the goal of the conventional few-shot video classification task (FSV) [48,2] is to learn a classifier which predicts labels among novel classes at test time. As the test-time label space is restricted to a few novel classes, the FSV setting is unrealistic. Thus, in

this paper, we additionally study the generalized few-shot video classification (GFSV) which allows videos at test time to belong to any base or novel class.

3.1 3D CNN for FSV (3DFSV)

In this section, we introduce our spatiotemporal CNN baseline for few-shot video classification (3DFSV). Our approach in Figure 2 consists of 1) a representation learning stage which trains a spatiotemporal CNN on the base classes, 2) a few-shot learning stage that trains a linear classifier for novel classes with few labeled videos, and 3) a testing stage which evaluates the model on unseen test videos. The details of each of these stages are given below.

Representation learning. Our model adopts a 3D CNN [40] $\phi : \mathbb{R}^{F\times 3\times H\times W} \rightarrow \mathbb{R}^{d_v}$, encoding a short, fixed-length video clip of F RGB frames with spatial resolution $H \times W$ to a feature vector in a d_v-dimensional embedding space. On top of the feature extractor ϕ, we define a linear classifier $f(\bullet; W_b)$ parameterized by a weight matrix $W_b \in \mathbb{R}^{d_v\times|\mathbf{C}_b|}$, producing a probability distribution over the base classes. The objective is to jointly learn the network ϕ and the classifier W_b by minimizing the cross-entropy classification loss on video clips randomly sampled from training videos $\mathbf{X}_b$ of base classes. More specifically, given a training video $\mathbf{x} \in \mathbf{X}_b$ with a label $\mathbf{y} \in \mathbf{C}_b$, the loss for a video clip $\mathbf{x}_i \in \mathbb{R}^{F\times 3\times H\times W}$ sampled from video $\mathbf{x}$ is defined as

$$\mathcal{L}(\mathbf{x}_i) = -\log \sigma(W_b^T \phi(\mathbf{x}_i))_{\mathbf{y}} \tag{1}$$

where σ denotes the softmax function that produces a probability distribution and $\sigma(\bullet)_{\mathbf{y}}$ is the probability at class $\mathbf{y}$. Following [4], we do not do meta-learning, so we can use all the base classes to learn the network ϕ.

Few-shot learning. This stage aims to adapt the learned network ϕ to recognize novel classes $\mathbf{C}_n$ with a few training videos $\mathbf{X}_n$. To reduce overfitting, we fix the network ϕ and learn a linear classifier $f(\bullet, W_n)$ by minimizing the cross-entropy loss on video clips randomly sampled from videos in $\mathbf{X}_n$, where $W_n \in \mathbb{R}^{d_v\times|\mathbf{C}_n|}$ is the weight matrix of the linear classifier. Similarly, we define the loss for a video clip $\mathbf{x}_i$ sampled from $\mathbf{x} \in \mathbf{X}_n$ with a label $\mathbf{y}$ as

$$\mathcal{L}(\mathbf{x}_i) = -\log \sigma(W_n^T \phi(\mathbf{x}_i))_{\mathbf{y}} \tag{2}$$

Testing. The spatiotemporal CNN operates on fixed-length video *clips* of F RGB frames and the classifiers make clip-level predictions. At test time, the model must predict the label of a test video $\mathbf{x} \in \mathbb{R}^{T\times 3\times H\times W}$ with arbitrary time length T. We achieve this by randomly drawing a set L of clips $\{\mathbf{x}_i\}_{i=1}^{L}$ from video $\mathbf{x}$, where $\mathbf{x}_i \in \mathbb{R}^{F\times 3\times H\times W}$. The video-level prediction is then obtained by averaging the prediction scores after the softmax function over those L clips. For few-shot video classification (FSV), this is:

$$\frac{1}{L}\sum_{i=1}^{L} f(\mathbf{x}_i; W_n). \tag{3}$$

For generalized few-shot video classification (GFSV), both base and novel classes are taken into account and we concatenate the base class weight W_b learned in the representation stage with the novel class weight W_n learned in the few-shot learning stage:

$$\frac{1}{L}\sum_{i=1}^{L} f(\mathbf{x}_i; [W_b; W_n]). \tag{4}$$

3.2 Retrieval-enhanced 3DFSV (R-3DFSV)

During few-shot learning, fine-tuning the network ϕ or learning the classifier $f(\bullet; W_n)$ alone is prone to overfitting. Moreover, class-labeled videos to be used for fine-tuning are scarce. Instead, the hypothesis is that leveraging a massive collection of weakly-labeled real-world videos would improve our novel-class classifier. To this end, for each novel class, we propose to retrieve a subset of weakly-labeled videos, associate pseudo-labels to these retrieved videos and use them to expand the training set of novel classes. It is worth noting that those retrieved videos may be assigned with wrong labels and have domain shift issues as they belong to another heterogeneous dataset, making this idea challenging to implement. For efficiency and to reduce the label noise, we adopt the following two-step retrieval approach.

Tag-based video retrieval. The YFCC100M dataset [36] includes around 800K videos collected from Flickr, with a total length of over 8000 hours. Processing a large collection of videos has a high computational demand and a large portion of them are irrelevant to our target classes. Thus, we restrict ourselves to videos with tags related to those of the target class names and leverage information that is complementary to the actual video content to increase the visual diversity.

Given a video with user tags $\{t_i\}_{i=1}^{S}$ where $t_i \in \mathcal{T}$ is a word or phrase and S is the number of tags, we represent it with an average tag embedding $\frac{1}{S}\sum_{i=1}^{S}\varphi(t_i)$. The tag embedding $\varphi(.) : \mathcal{T} \rightarrow \mathbb{R}^{d_t}$ maps each tag to a d_t dimensional embedding space, e.g., Fasttext [18]. Similarly, we can represent each class by the text embedding of its class name and then for each novel class c, we compute its cosine similarity to all the video tags and retrieve the N most similar videos according to this distance.

Selecting best clips. The video tag retrieval selects a list of N candidate videos for each novel class. However, those videos are not yet suitable for training because the annotation may be erroneous, which can harm the performance. Besides, some weakly-labeled videos can last as long as an hour. We thus propose to select the best short clips of F frames from those candidate videos using the few-shot videos of novel classes.

Given a set of few-shot videos $\mathbf{X}_n^c$ from novel class c, we randomly sample L video clips from each video. We then extract features from those clips with the spatiotemporal CNN ϕ and compute the class prototype by averaging over clip features. Similarly, for each retrieved candidate video of novel class c, we also randomly draw L video clips and extract clip features from ϕ. Finally, we perform a nearest neighbour search with cosine distance to find the M best matching clips of the class prototype:

$$\max_{\mathbf{x}_j} \cos(p_c, \phi(\mathbf{x}_j)) \tag{5}$$

where p_c denotes the class prototype of class c, $\mathbf{x}_j$ is the clip belonging to the retrieved weakly-labeled videos. After repeating this process for each novel class, we obtain a collection of pseudo-labeled video clips $\mathbf{X}_p = \{\mathbf{X}_p^c\}_{c=1}^{|C_n|}$ where $\mathbf{X}_p^c$ indicates the best M video clips from YFCC100M for novel class c.

Batch denoising. The retrieved video clips contribute to learning a better novel-class classifier $f(\bullet; W_n)$ in the few-shot learning stage by expanding the training set of novel classes from $\mathbf{X}_n$ to $\mathbf{X}_n \bigcup \mathbf{X}_p$. $\mathbf{X}_p$ may include video clips with wrong labels. During the optimization, we adopt a simple strategy to alleviate the noise: half of the video clips per batch come from $\mathbf{X}_n$ and another half from $\mathbf{X}_p$ at each iteration. The purpose is to reduce the gradient noise in each mini-batch by enforcing that half of the samples are trustworthy.

4 Experiments

In this section, we first describe the existing experimental settings and our proposed setting for few-shot video recognition. We then present the results comparing our approaches with the state-of-the-art methods in the existing setting on two datasets, the results of our approach in our proposed settings, model analysis and qualitative results.

4.1 Experimental settings

Here we describe the four datasets we use, previous few-shot video classification protocols and our settings.

Datasets. Kinetics [22] is a large-scale video classification dataset which covers 400 human action classes including human-object and human-human interactions. Its videos are collected from Youtube and trimmed to include only one action class. The UCF101 [35] dataset is also collected from Youtube videos, consisting of 101 realistic human action classes, with one action label in each video. SomethingV2 [14] is a fine-grained human action recognition dataset, containing 174 action classes, in which each video shows a human performing a predefined basic action, such as "picking something up" and "pulling something from left to right". We use the second release of the dataset. YFCC100M [36] is the largest publicly available multimedia collection with about 99.2 million images and 800k videos from Flickr. Although none of these videos are annotated with a class label, half of them (400k) have at least one user tag. We use the tag-labeled videos of YFCC100M to improve the few-shot video classification.

Prior setup. The existing practice of [48] and [2] indicates randomly selecting 100 classes on Kinetics and on SomethingV2 datasets respectively. Those 100 classes are then randomly divided into 64, 12, and 24 non-overlapping classes to construct the meta-training, meta-validation and meta-testing sets. The meta-training and meta-validation sets are used for training models and tuning hyperparameters. In the testing phase of this meta-learning setting [48,2], each episode simulates a n-way, k-shot classification problem by randomly sampling a support set consisting of k samples from each of the n classes, and a query set consisting of one sample from each of the n classes. While the support set is used to adapt the model to recognize novel classes, the classification

	# classes			# videos		
	train	**val**	**test**	**train**	**val**	**test**
Kinetics	64	12	24	6400	1200	2400+2288
UCF101	64	12	24	5891	443	971+1162
SomethingV2	64	12	24	67013	1926	2857+5243

Table 1. Statistics of our data splits on Kinetics, UCF101 and SomethingV2 datasets. We follow the train, val, and test class splits of [48] and [2] on Kinetics and SomethingV2 respectively. In addition, we add test videos (the second number under the second test column) from train classes for GFSV. We also introduce a new data split on UCF101 and for all datasets we propose 5-,10-,15-,24-way (the maximum number of test classes) and 1-,5-shot setting.

accuracy is computed at each episode on the query set and mean top-1 accuracy over 20,000 episodes constitutes the final accuracy.

Proposed setup. The prior experimental setup is limited to $n = 5$ classes in each episode, even though there are 24 novel classes in the test set. As in this setting the performance saturates quickly, we extend it to 10-way, 15-way and 24-way settings. Similarly, the previous meta-learning setup assumes that test videos all come from novel classes. On the other hand, it is important in many real-world scenarios that the classifier does not forget about previously learned classes while learning novel classes. Thus, we propose the more challenging generalized few-shot video classification (GFSV) setting where the model needs to predict both base and novel classes.

To evaluate a n-way k-shot problem in GFSV, in addition to a support and a query set of novel classes, at each test episode we randomly draw an additional query set of 5 samples from each of the 64 base classes. We do not sample a support set for base classes because base class classifiers have been learned during the representation learning phase. We report the mean top-1 accuracy of both base and novel classes over 500 episodes.

Kinetics, UCF101 and SomethingV2 datasets are used as our few-shot video classification datasets with disjoint sets of train, validation and test classes (see Table 1 for details). Here we refer to base classes as train classes. Test classes include the classes we sample novel classes from in each testing episode. For Kinetics and SomethingV2, we follow the splits proposed by [48] and [2] respectively for a fair comparison. It is worth noting that 3 out of 24 test classes in Kinetics appear in Sports1M, which is used for pretraining our 3D ConvNet. But the performance drop is negligible if we replace those 3 classes with other 3 random kinetics classes that are not present in Sports1M (more details can be found in the supplementary material). Following the same convention, we randomly select 64, 12 and 24 non-overlapping classes as train, validation and test classes from UCF101 dataset, which is widely used for video action recognition. We ensure that in our splits the novel classes do not overlap with the classes of Sports1M. For the GFSV setting, in each dataset the test set includes samples from base classes coming from the validation split of the original dataset.

	Kinetics		SomethingV2	
Method	**1-shot**	**5-shot**	**1-shot**	**5-shot**
CMN [48]	60.5	78.9	-	-
CMN++ [2]	65.4	78.8	34.4	43.8
TAM [2]	73.0	85.8	42.8	52.3
3DFSV (ours, scratch)	48.9	67.8	57.9	75.0
3DFSV (ours, pretrained)	92.5	**97.8**	**59.1**	**80.1**
R-3DFSV (ours, pretrained)	**95.3**	**97.8**	-	-

Table 2. Comparing with the state-of-the-art few-shot video classification methods. We report top-1 accuracy on the novel classes of Kinetics and SomethingV2 for 1-shot and 5-shot tasks (both in 5-way). 3DFSV (ours, scratch): our R(2+1)D is trained from scratch; 3DFSV (ours, pretrained): our model is trained from the Sports1M-pretrained R(2+1)D. R-3DFSV (ours, pretrained): our model with retrieved videos, trained from the Sports1M-pretrained R(2+1)D.

Implementation details. Unless otherwise stated our backbone is a 34-layer R(2+1)D [40] pretrained on Sports1M [21] which takes as input video clips consisting of $F = 16$ RGB frames with spatial resolution of $H = 112 \times W = 112$. We extract clip features from the $d_v = 512$ dimensional top pooling units of the R(2+1)D.

In the representation learning stage, we fine-tune the R(2+1)D with a constant learning rate 0.001 on all datasets and stop training when the validation accuracy of base classes saturates. We perform standard spatial data augmentation including random cropping and horizontal flipping. We also apply temporal data augmentation by randomly drawing 8 clips from a video in one epoch. In the few-shot learning stage, the same data augmentation is applied and the novel class classifier is learned with a constant learning rate 0.01 for 10 epochs on all the datasets. At test time, we randomly draw $L = 10$ clips from each video and average their predictions for a video-level prediction.

As for the retrieval approach, we use the 400 dimensional ($d_t = 400$) fasttext [17] embedding trained with GoogleNews. We first retrieve $N = 20$ candidate videos for each class with video tag retrieval and then select $M = 5$ best clips among those videos with visual similarities.

4.2 Comparing with the state-of-the-art

In this section, we compare our model with the state-of-the-art in existing evaluation settings which mainly consider 1-shot, 5-way and 5-shot, 5-way problems and evaluate only on novel classes, i.e., FSV. The baselines CMN [48] and TAM [2] are considered as the state-of-the-art in few-shot video classification. CMN [48] proposes a multi-saliency embedding function to extract video descriptor, and few-shot classification is then done by the compound memory network [20]. TAM [2] proposes to leverage the long-range temporal ordering information in video data through temporal alignment. They additionally build a stronger CMN, namely CMN++, by using the few-shot learning practices from [4]. We use their reported numbers for fair comparison. The results are shown in Table 2. As the code from CMN [48] and TAM [2] is not available at the time of submission we do not include UCF101 results.

On Kinetics, we observe that our 3DFSV (pretrain) approach, i.e. without retrieval, outperforms the previous best results by over 19% in 1-shot case (73.0% of TAM vs 92.5% of ours), and by 12% in 5-shot case ($85.8.0\%$ of TAM vs 97.8% of ours). On SomethingV2 dataset, we would like to first highlight that our 3DFSV (scratch) significantly improves over TAM by 15.1% in 1-shot (42.8% of TAM vs 57.9% of ours) and by surprisingly 22.7% in 5-shot (52.3% of TAM vs 75.0% of ours). This is encouraging because the 2D CNN backbone of TAM is pretrained on ImageNet, while our R(2+1)D backbone is trained from random initialization.

Our 3DFSV (pretrain) yields further improvement after using the Sports1M-pretrained R(2+1)D. We observe that the effect of the Sports1M-pretrained model on SomethingV2 is not as significant as on Kinetics because there is a large domain gap between Sports1M to SomethingV2 datasets. Those results show that a simple linear classifier on top of a pretrained 3D CNN, e.g. R(2+1)D [40], performs better than sophisticated methods with a pretrained 2D ConvNet as a backbone.

Although as shown in C3D [38], I3D [3], R(2+1)D [40], spatiotemporal CNNs have an edge over 2D spatial ConvNet [16] in the fully supervised video classification with enough annotated training data, we are the first to apply R(2+1)D in the few-shot video classification with limited labeled data. It is worth noting that our R(2+1)D is pretrained on the Sports1M while the 2D ResNet backbone of CMN [48] and TAM [2] is pretrained on ImageNet. A direct comparison between 3D CNNs and 2D CNNs is hard because they are designed for different input data. While it is standard to use an ImageNet-pretrained 2D CNN in image domains, it is common to apply a Sports1M-pretrained 3D CNN in video domains. One of our goals is to establish a strong few-shot video classification baseline with 3D CNNs. Intuitively, the temporal cue of the video is better preserved when clips are processed directly by a spatiotemporal CNN as opposed to processing them as images via a 2D ConvNet. Indeed, even though we train our 3DFSV from the random initialization on SomethingV2 dataset which requires strong temporal information, our results still remain promising. This confirms the importance of 3D CNNs for few-shot video classification.

Our R-3DFSV (pretrain) approach, i.e. with retrieved weakly-labeled video clips, lead to further improvements in 1-shot case (3DFSV (pretrain) 92.5% vs R-3DFSV (pretrain) 95.3) on Kinetics dataset. This implies that weakly-labeled videos retrieved from the YFCC100M dataset include discriminative cues for Kinetics tasks. In 5-shot, our R-3DFSV (pretrain) approach achieves similar performance as our 3DFSV (pretrain) approach however with an 97.8% this task is almost saturated. We do not retrieve any weakly-labeled videos for the SomethingV2 dataset because it is a fine-grained dataset of basic actions and it is unlikely that YFCC100M includes any relevant video for that dataset. As a summary, although 5-way classification setting is still challenging to those methods with 2D ConvNet backbone, the results saturate with the stronger spatiotemporal CNN backbone.

4.3 Increasing the number of classes in FSV

Although prior works evaluated few-shot video classification on 5-way, i.e. the number of novel classes at test time is 5, our 5-way results are already saturated. Hence, in this section, we go beyond 5-way classification and extensively evaluate our approach in the

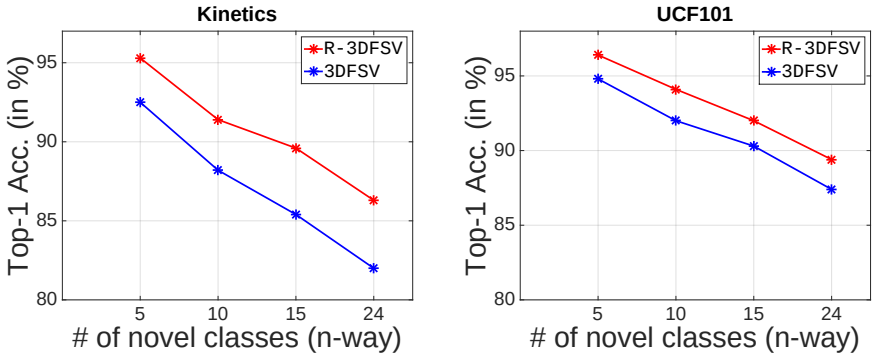

Fig. 3. Results of 3DFSV and R-3DFSV on both Kinetics and UCF101 in the one-shot video classification setting (FSV). In this experiment we go beyond the classical 5-way classification setting. We use 5, 10, 15 and 24 (all) of the novel classes in each testing episode. We report the top-1 accuracy of novel classes.

more challenging, i.e., 10-way, 15-way and 24-way few-shot video classification (FSV) setting. Note that from every class we use one sample per class during training, i.e. one-shot video classification.

As shown in Figure 3, our R-3DFSV method exceeds 95% accuracy both in Kinetics and UCF101 datasets for 5-way classification. With the increasing number of novel classes, e.g. 10, 15 and 24, as expected, the performance degrades. Note that, our R-3DFSV approach with retrieval consistently outperforms our 3DFSV approach without retrieval and the more challenging the task becomes, e.g. from 5-way to 24-way, the larger improvement retrieval approach can achieve on Kinetics, i.e. our retrieval-based method is better than our baseline method by 2.8% in 5-way (ours 3DFSV 92.5% vs our R-3DFSV 95.3%) and the gap becomes 4.3% in 24-way (our 3DFSV 82.0% vs our R-3DFSV 86.3%).

The trend with a decreasing accuracy by going from 5-way to 24-way indicates that the more realistic task on few-shot video classification has not yet been solved even with a spatiotemporal CNN. We hope that these results will encourage more progress in this challenging setting of many-way few-shot video classification setting.

4.4 Evaluating base and novel classes in GFSV

The FSV setting has a strong assumption that test videos all come from novel classes. In contrast to the FSV, GFSV is more realistic and requires models to predict both base and novel classes in each testing episode. In other words, 64 base classes become distracting classes when predicting novel classes which makes the task more challenging. Intuitively, distinguishing novel and base classes is a challenging task because there are severe imbalance issues between the base classes with a large number of training examples and the novel classes with only few-shot examples. In this section, we evaluate our methods in the more realistic and challenging generalized few-shot video classification (GFSV) setting.

	Method	Kinetics novel	Kinetics base	UCF101 novel	UCF101 base
1-shot	3DFSV	7.5	88.7	3.5	97.1
	R-3DFSV	13.7	88.7	4.9	97.1
5-shot	3DFSV	20.5	88.7	10.1	97.1
	R-3DFSV	22.3	88.7	10.4	97.1

Table 3. Generalized few-shot video classification (GFSV) results on Kinetics and UCF101 in 5-way 1-shot and 5-shot tasks. We report top-1 accuracy on both base and novel classes.

PR	SS	RL	VR	BD	BC	Acc
✓						27.1
		✓				48.9
	✓	✓				51.9
✓		✓				92.5
✓		✓	✓			91.4
✓		✓	✓	✓		93.2
✓		✓	✓	✓	✓	95.3

Table 4. 5-way 1-shot results ablated on Kinetics. **PR**: pretraining on Sports1M; **SS**: self-supervised pretraining [23]; **RL**: representation learning on base classes; **VR**: retrieve videos with tags [36]; **BD**: batch denoising. **BC**: best clip selection.

In Table 3, on the Kinetics dataset, we observe a large performance gap between base and novel classes in both 1-shot and 5-shot cases, i.e., 3DFSV only achieves 7.5% on novel classes vs 88.7% on base classes. The reason is that predictions of novel classes are dominated by the base classes. Interestingly, our R-3DFSV improves 3DFSV on novel classes in both 1-shot and 5-shot cases, e.g., 7.5% of 3DFSV vs 13.7% of R-3DFSV in 1-shot. A similar trend can be observed on the UCF101 dataset. Those results demonstrate that our retrieval-based approach can alleviate the imbalance issues to some extent. At the same time, we find that generalized few-shot video classification (GFSV) setting, e.g. not restricting the test time search space only to novel classes but considering all of the classes even though base classes are distracting, is still a challenging task and hope that this setting will attract interest of a wider community for future research.

4.5 Ablation study and retrieved clips

In this section, we perform an ablation study to understand the importance of each component of our approach. After the ablation study, we evaluate the importance of the number of retrieved clips to the few-shot video classification (FSV) performance.

Ablation study. We ablate our model in the 1-shot, 5-way video classification task on Kinetics dataset with respect to six critical parts including pretraining R(2+1)D on Sports1M (**PR**), self-supervised model of [23] as the backbone (**SS**), representation learning on base classes (**RL**), video retrieval with tags (**VR**), batch denoising (**BD**) and best clip selection (**BC**). Table 4 shows the results.

We start from a model with only a few-shot learning stage on novel classes. If a **PR** component is added to the model (first result row in Table (4), the newly-obtained model can achieve 27.1% accuracy which is only slightly better than random guessing performance (20%). It demonstrates that a pretrained 3D CNN alone is not sufficient for a good performance. Besides, it also indicates that there exists a domain shift between the pretraining dataset, i.e. Sports1M, and our target Kinetics dataset.

Adding **RL** component to the model (the second result row) means to train representation on base classes from scratch, which results in a worse accuracy of 48.9%

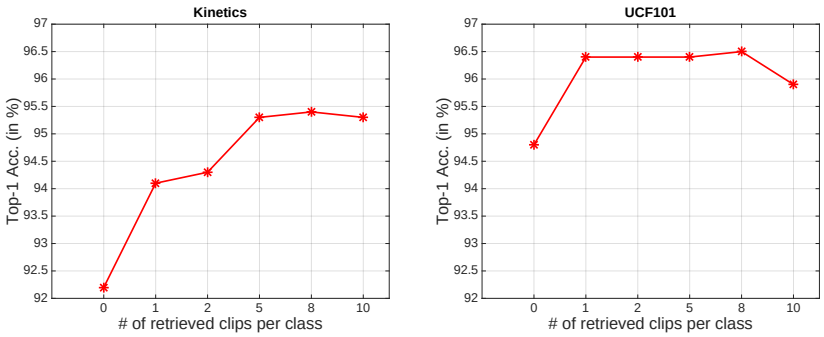

Fig. 4. The effect of increasing the number of retrieved clips, **left**: on Kinetics, **right**: on UCF101. Both experiments are conducted on the one-shot, five-way classification task, reporting top-1 accuracy in the few-shot video classification (FSV) setting.

compared to our full model. The primary reason for worse results is that optimizing the massive number of parameters of R(2+1)D is difficult on a train set consisting of only 6400 videos. Interestingly, if we adopt the self-supervised pretrained 3D CNN (MC3 pretrained on Kinetics without using any label) of [23], i.e., **SS**, we immediate get 3.0% performance gains (the third result row) over training from random initialization. Adding both **PR** and **RL** components (the fourth row) obtains an accuracy of 92.5 which significantly boosts adding **PR** and **RL** components alone.

Next, we study two critical components proposed in our retrieval approach. Comparing to our approach without retrieval (the fourth row), directly appending retrieved videos from YFCC100M (**VR**) to the few-shot training set of novel classes (the fifth result row) leads to 0.9% performance drop, while performing the batch denoising (the sixth row) in addition to **VR** obtains 0.7% gain. This implies that noisy labels from retrieved videos may hurt the performance but our batch denoising technique handles the noise well. Finally, adding the best clip selection (**BC**, the last row) after **VR** and **BD** gets a big boost of 2.8% accuracy. In summary, those ablation studies demonstrate the effectiveness of the six different critical parts in our approach.

Influence of the number of retrieved clips. Intuitively, when the number of retrieved clips increases, the retrieved videos become more diverse, but at the same time, the risk of obtaining negative videos becomes higher. We show the effectiveness of our R-3DFSV with the increasing number of retrieved clips in Figure 4.

On the Kinetics dataset (left of Figure 4), without retrieving any videos, the performance is 92.5%. As we increase the number of retrieved video clips for each novel class, the performance keeps improving and saturates at retrieving 8 clips per class, reaching an accuracy of 95.4%. On the UCF101 dataset (right of Figure 4), retrieving 1 clip gives us 1.6% gain. Retrieving more clips does not further improve the results, indicating more negative videos are retrieved. On the other hand, our batch denoising strategy is able to tolerate the noise to some extent. We observe a slight performance drop at retrieving 10 clips because the noise level becomes too high, i.e. there are 10 times more noisy labels than clean labels.

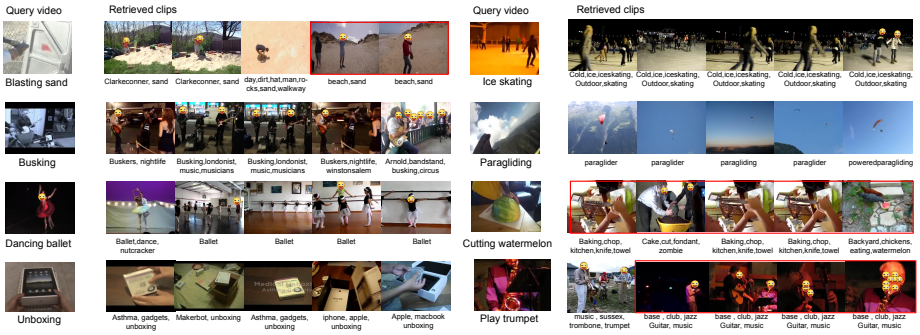

Fig. 5. Top-5 retrieved video clips from YFCC100M for 8 novel classes on Kinetics. The left column is the class name with its one-shot query video and the right column shows the retrieved 16-frame video clips (middle frame is visualized) together with their users tags. Negative retrievals are marked in red.

4.6 Qualitative results

In Figure 5, we visualize the top-5 video clips we retrieve from YFCC100M dataset with video tag retrieval followed by the best clips selection. Here we only show 8 novel classes of Kinetics dataset due to the space limitation and visualization of other classes are in supplementary.

We observe that the retrieved video clips of some classes are of high quality, meaning that those videos truly reveal the target novel classes. For instance, retrieved clips of class "Busking" are all correct because user tags of those videos consist of words like "buskers", "busking" that are close to the class name, and the best clip selection can effectively filter out the irrelevant clips. It is intuitive those clips can potentially help to learn better novel class classifiers by supplementing the limited training videos.

Failure cases are also common. For example, videos from the class "Cutting watermelon" do not retrieve any positive videos. The reasons can be that there are no user tags of cutting watermelon or our tag embeddings are not good enough. Those negative videos might hurt the performance if we treat them equally, which is why the batch denoising is critical to reduce the effect of negative videos.

5 Conclusion

In this work, we point out that a spatiotemporal CNN trained on a large-scale video dataset saturates existing few-shot video classification benchmarks. Hence, we propose new more challenging experimental settings, namely generalized few-shot video classification (GFSV) and few-shot video classification with more ways than the classical 5-way setting. We further improve spatiotemporal CNNs by leveraging the weakly-labeled videos from YFCC100M using weak-labels such as tags for text-supported and video-based retrieval. Our results show that generalized more-way few-shot video classification is challenging and we encourage future research in this setting.

References

1. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2016) 3
2. Cao, K., Ji, J., Cao, Z., Chang, C.Y., Niebles, J.C.: Few-shot video classification via temporal alignment. In: CVPR (2020) 1, 2, 4, 7, 8, 9, 10
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017) 10
4. Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.C.F., Huang, J.B.: A closer look at few-shot classification. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=HkxLXnAcFQ 1, 3, 5, 9
5. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: Proceedings of the 14th International Conference on Computer Communications and Networks. VS-PETS Beijing, China (2005) 4
6. Douze, M., Szlam, A., Hariharan, B., Jégou, H.: Low-shot learning with large-scale diffusion. In: CVPR (2018) 3
7. Dwivedi, S.K., Gupta, V., Mitra, R., Ahmed, S., Jain, A.: Protogan: Towards few shot learning for action recognition. arXiv preprint arXiv:1909.07945 (2019) 3
8. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6202–6211 (2019) 4
9. Feichtenhofer, C., Pinz, A., Wildes, R.: Spatiotemporal residual networks for video action recognition. In: Advances in neural information processing systems. pp. 3468–3476 (2016) 4
10. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1933–1941 (2016) 4
11. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1126–1135. JMLR. org (2017) 3
12. Ghadiyaram, D., Tran, D., Mahajan, D.: Large-scale weakly-supervised pre-training for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12046–12055 (2019) 4
13. Girdhar, R., Ramanan, D., Gupta, A., Sivic, J., Russell, B.: Actionvlad: Learning spatio-temporal aggregation for action classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 971–980 (2017) 4
14. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: ICCV (2017) 7
15. Hariharan, B., Girshick, R.: Low-shot visual recognition by shrinking and hallucinating features. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3018–3027 (2017) 2, 3
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 10
17. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: Fasttext. zip: Compressing text classification models. arXiv preprint arXiv:1612.03651 (2016) 9
18. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 427–431. Association for Computational Linguistics (April 2017) 6

19. Joulin, A., van der Maaten, L., Jabri, A., Vasilache, N.: Learning visual features from large weakly supervised data. In: European Conference on Computer Vision. pp. 67–84. Springer (2016) 3
20. Kaiser, Ł., Nachum, O., Roy, A., Bengio, S.: Learning to remember rare events. arXiv preprint arXiv:1703.03129 (2017) 9
21. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR (2014) 4, 9
22. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017) 4, 7
23. Korbar, B., Tran, D., Torresani, L.: Cooperative learning of audio and video models from self-supervised synchronization. In: NeurIPS (2018) 4, 12, 13
24. Laptev, I.: On space-time interest points. International journal of computer vision **64**(2-3), 107–123 (2005) 4
25. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., van der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 181–196 (2018) 4
26. Mensink, T., Verbeek, J., Perronnin, F., Csurka, G.: Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In: ECCV (2012) 2
27. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 631–648 (2018) 4
28. Qi, H., Brown, M., Lowe, D.G.: Low-shot learning with imprinted weights. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5822–5830 (2018) 1, 3
29. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: ICLR (2016) 1, 2, 3
30. Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J.B., Larochelle, H., Zemel, R.S.: Meta-learning for semi-supervised few-shot classification. In: ICLR (2018) 3
31. Sadanand, S., Corso, J.J.: Action bank: A high-level representation of activity in video. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1234–1241. IEEE (2012) 4
32. Schoenfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., Akata, Z.: Generalized zero- and few-shot learning via aligned variational autoencoders. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3
33. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems. pp. 568–576 (2014) 4
34. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: NeurIPS (2017) 3
35. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012) 7
36. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. arXiv preprint arXiv:1503.01817 (2015) 1, 3, 4, 6, 7, 12
37. Torralba, A., Efros, A.A., et al.: Unbiased look at dataset bias. In: CVPR (2011) 3
38. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: CVPR (2015) 4, 10
39. Tran, D., Wang, H., Torresani, L., Feiszli, M.: Video classification with channel-separated convolutional networks. arXiv preprint arXiv:1904.02811 (2019) 4
40. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: CVPR (2018) 4, 5, 9, 10

41. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: Advances in neural information processing systems. pp. 3630–3638 (2016) 1, 3
42. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the IEEE international conference on computer vision. pp. 3551–3558 (2013) 4
43. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. pp. 20–36. Springer (2016) 4
44. Wang, Y., Chao, W.L., Weinberger, K.Q., van der Maaten, L.: Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. arXiv preprint arXiv:1911.04623 (2019) 3
45. Xian, Y., Sharma, S., Schiele, B., Akata, Z.: F-vaegan-d2: A feature generating framework for any-shot learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3
46. Yalniz, I.Z., Jégou, H., Chen, K., Paluri, M., Mahajan, D.: Billion-scale semi-supervised learning for image classification. arXiv preprint arXiv:1905.00546 (2019) 4
47. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4694–4702 (2015) 4
48. Zhu, L., Yang, Y.: Compound memory networks for few-shot video classification. In: ECCV (2018) 1, 2, 3, 4, 7, 8, 9, 10