# Few-Shot Ensemble Learning for Video Classification with SlowFast Memory Networks

Mengshi Qi*
CVLab, EPFL
Switzerland

Jie Qin*†
Inception Institute of
Artificial Intelligence (IIAI), UAE

Xiantong Zhen
IIAI, UAE
Universiteit van Amsterdam, NL

Di Huang
Beijing Advanced Innovation Center
for BDBC, Beihang University
Beijing, China

Yi Yang
ReLER Lab
University of Technology Sydney
Sydney, Australia

Jiebo Luo
Department of Computer Science
University of Rochester
NY, USA

## ABSTRACT

In the era of *big data*, few-shot learning has recently received much attention in multimedia analysis and computer vision due to its appealing ability of learning from scarce labeled data. However, it has been largely underdeveloped in the video domain, which is even more challenging due to the huge spatial-temporal variability of video data. In this paper, we address few-shot video classification by learning an ensemble of SlowFast networks augmented with memory units. Specifically, we introduce a family of few-shot learners based on SlowFast networks which are used to extract informative features at multiple rates, and we incorporate a memory unit into each network to enable encoding and retrieving crucial information instantly. Furthermore, we propose a choice controller network to leverage the diversity of few-shot learners by learning to adaptively assign a confidence score to each SlowFast memory network, leading to a strong classifier for enhanced prediction. Experimental results on two widely-adopted video datasets demonstrate the effectiveness of the proposed method, as well as its superior performance over the state-of-the-art approaches.

## CCS CONCEPTS

• **Computing methodologies** → **Ensemble methods**; **Activity recognition and understanding**.

## KEYWORDS

Ensemble Learning, Few-Shot Learning, Video Classification, Memory Network

**ACM Reference Format:**
Mengshi Qi, Jie Qin, Xiantong Zhen, Di Huang, Yi Yang, and Jiebo Luo. 2020. Few-Shot Ensemble Learning for Video Classification with SlowFast Memory Networks. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3416269

*Equal contribution.
†Corresponding author: Jie Qin (qinjiebuaa@gmail.com).


*MM '20, October 12–16, 2020, Seattle, WA, USA*

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3416269

## 1 INTRODUCTION

With the emergence of social networks and short-video sharing applications (*e.g., Instagram* and *TikTok*), there is an explosive growth of multimedia data (especially videos) on the Internet. Since manually annotating each video is infeasible, how to recognize unseen/novel videos becomes a very challenging problem in multimedia interpretation and understanding. To this end, few-shot video classification has emerged as a novel task [2, 52, 53, 56], aiming to learn a classifier to recognize novel classes given only a few labeled video examples. Recently, the task has received increasing attention due to its great potential of use in diverse applications, *e.g.,* media content understanding, social media analysis, human-computer interaction, and intelligent surveillance [4, 5, 13, 17, 21, 22, 24, 27–34, 39, 50, 53–55, 57]. However, the limited number of annotated examples per class cannot well represent the overall class distribution, making this task extremely challenging and worthy of further research efforts.

So far, extensive efforts have been devoted to few-shot learning, which can be mainly divided into meta-learning based approaches [10, 35, 36] and metric learning based ones [14, 25, 42, 43, 46]. The former solves the problem by designing a meta-learner to learn a base-learner that is able to effectively and efficiently adapt to unseen related tasks. The latter follows a 'learning to compare' paradigm that classifies an unseen image by measuring the similarity based on a certain distance metric between two images, leading to a deep embedding space for knowledge transfer. Nevertheless, most current studies focus on few-shot learning in the image domain, and there are very limited works addressing few-shot video classification. In contrast to images, video data is usually in a much higher dimensional space of more complex structures in terms of both spatial and temporal variations, which makes few-shot video classification even more challenging, especially when there is a shortage of labeled data.

In particular, there are two outstanding issues to solve in few-shot video classification. First, since the labeled videos are very limited and videos are of complex spatial-temporal structures, more informative and discriminative video representations need to be extracted. Second, a single base-learner can be too weak to handle

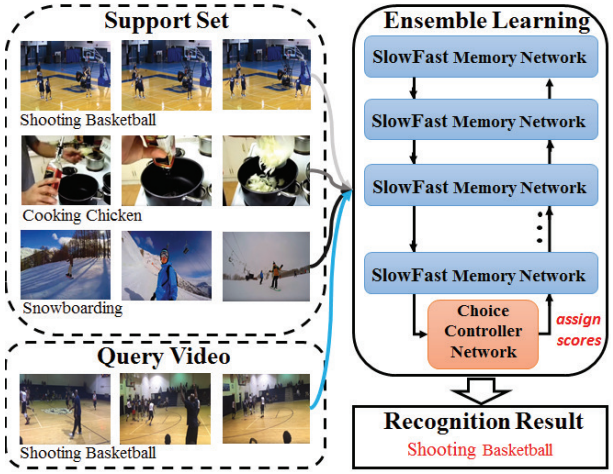


**Figure 1: Illustration of 3-way 1-shot few-shot video classification and our proposed ensemble learning model. Given a support set that contains three categories with one video example per category, an ensemble learning architecture consisted of a collection of SlowFast memory networks is trained for meta-learning following the episode training strategy, and evaluated for the generalization performance on new classes during test. In particular, a choice controller network in the proposed ensemble model is designed to generate various confidence scores to predict the final result.**

the huge spatial-temporal variability of video data with inadequate training samples, and the performance and robustness would degrade significantly when given new data during test. Therefore, it is highly desirable to build a strong learner from a collection of weak learners. However, how to exploit the diversity and capture the relevance between multiple few-shot learners remains an open and challenging problem.

In this work, we address the above issues by proposing a novel ensemble learning architecture to aggregate a set of weak learners for few-shot video classification. As illustrated in Fig. 1, the proposed approach encourages the learners to cooperate with each other while preserving sufficient diversities during training. Specifically, a family of SlowFast networks are employed as the base learners to capture both fine-grained spatial details and temporal dynamics at multiple rates. A memory unit is incorporated into each of the SlowFast networks to instantly encode and retrieve the relevant information. As a result, the representation and memorization capabilities of our network are largely enhanced even given only very few video examples. Furthermore, we develop a choice controller network to help the model to make the best final decision by adaptively assigning various confidence scores to each SlowFast memory network, and devising a novel hinge loss function to alleviate the overconfidence challenge. This leads to a stronger and more robust learner based on the consensus of multiple weak learners. Finally, we adopt the common episode training strategy [46] to gradually accumulate the learned representations of videos into our proposed model. Our main contributions include:

**1)** We propose a novel ensemble learning framework for few-shot video classification, which explicitly addresses the critical issues existing in video representations and weak learners due to the lack of labeled data.

**2)** We design a family of SlowFast networks to capture informative and discriminative video representations at different frame rates, which augmented with the memory unit can efficiently record and retrieve essential information to meet the few-shot requirement.

**3)** We develop a choice controller network to build a strong classifier to improve the prediction, by adaptively generating an optimal combination of weaker learners based on their diversitie.

We conduct extensive experiments on two popular benchmark datasets. The experimental results demonstrate that the proposed approach can achieve high performance and significantly outperforms state-of-the-art methods. The ablation study further verifies the high effectiveness for few-shot video classification.

## 2 RELATED WORK

### 2.1 Few-Shot Recognition

Few-shot learning aims to recognize novel categories with very limited annotated data [2, 52, 53]. Current research on few-shot learning can be divided into two perspectives: meta-learning based and metric-learning based methods. Meta-learning based approaches [10, 35, 36] usually train a meta-learner to update the parameters of the learner. For example, an Long-Short Term Memory (LSTM) [16] based meta-learner [35] was proposed to optimize a neural network classifier. Santoro *et al.* [36] adopted LSTM as a controller with an external memory module. Finn *et al.* [10] introduced a model-agnostic approach that could learn adaptable parameters for deep networks. Differently, metric-learning based ones [14, 25, 42, 43, 46] usually adopt some informative similarity metrics. For instance, Vinyals *et al.* [46] presented a bidirectional LSTM based Matching Networks for one-shot learning. Garcia *et al.* [14] leveraged the few-shot learning problem as a supervised message passing task with graph neural networks. Prototypical Networks [42] learnt a metric space to compute distances between different classes. Relation Network (RN) [43] learnt a non-linear deep metric by generating images' feature embedding with episodic training. A Simple Neural AttentIve Learner (SNAIL) [25] was introduced to combine temporal convolutions and soft attention. However, all the above works addressed few-shot image recognition rather than recognizing video data. In this study, we propose an ensemble learning based model to especially handle few-shot video classification.

### 2.2 Video Classification

A number of early works employed various hand-crafted spatial-temporal features for video classification, *e.g.*, histograms of flows (HOF) [20], motion boundary histograms (MBH) [6] and trajectories [47]. With the emergence of deep learning, two-stream convolutional neural networks (CNN) [41], 3D CNN [45], recurrent neural networks (RNN) [51], and other works [3, 8, 13, 17, 48, 50] have achieved more promising performance. However, very limited works [56] have been devoted to few-shot video classification. Because it is too expensive to annotate a huge number of videos, the networks should be able to be trained to classify each unseen

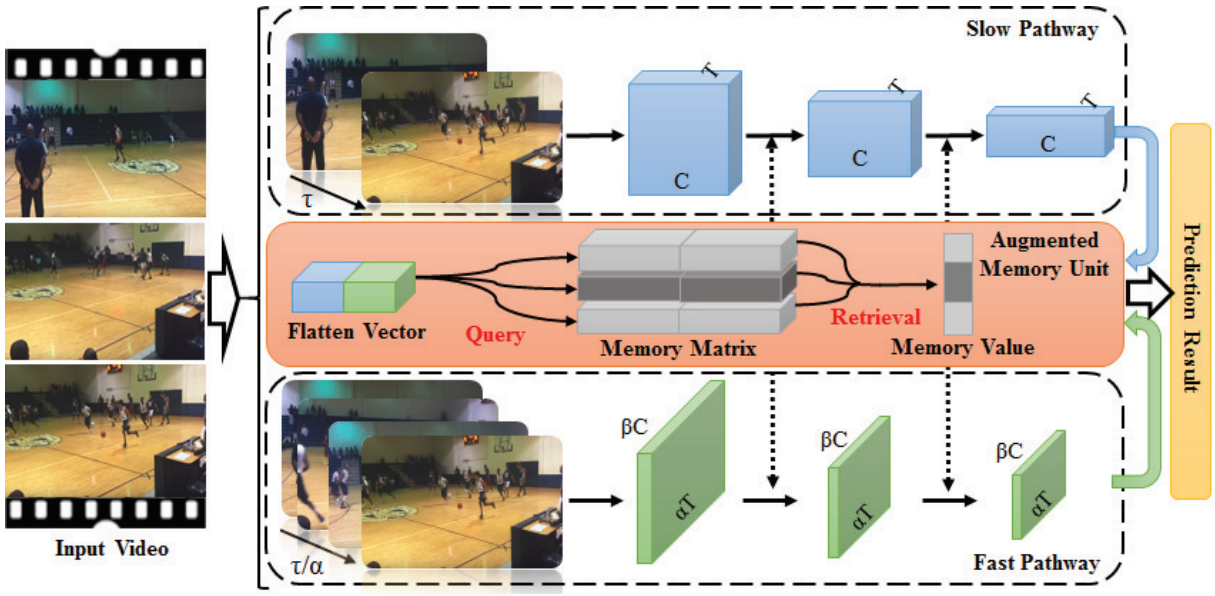


**Figure 2: Architecture of the proposed SlowFast memory network, which consists of a slow pathway at $\tau$ frame rate to extract low temporal resolution features, a fast pathway at $\tau/\alpha$ frame rate to capture higher temporal resolution representations with a fraction ($\beta$) of channels, and a memory unit designed to record and retrieve the most similar memory for classification. $C$ and $T$ denotes the number of channels and the temporal dimension of features, respectively. $\tau$ is the value of the temporal stride.**

category. In [56], compound memory networks were proposed to address the above video-based task; however, the basic idea was still derived from a few-shot image classification model. Conversely, we introduce an ensemble learning architecture with multi-rate SlowFast memory networks that specifically focuses on the task of few-shot video classification.

### 2.3 Ensemble Learning

In order to overcome the unreliability of a single model, ensemble learning [7, 11, 26, 44] aims at enhancing multiple weak learners by exploiting the diversity among them, resulting in a strong ensemble learner for better performance. Bagging [1] and boosting [12, 37, 38] are two common strategies. Meanwhile, traditional ensemble learning models were based on decision trees, random forests, and *etc.* Recently, more attention has been paid to CNN based ensemble learning. Due to the unsatisfactory performance of the single model, our goal is to study the effectiveness of ensemble learning specifically in the scenario of few-shot video classification. To the best of our knowledge, this is the first model of ensemble learning for video-based few-shot learning.

## 3 PROBLEM DEFINITION

In this work, we aim to train a model that can be utilized to classify unseen classes with a few training examples. Generally, we follow the meta-learning setting used in few-shot image classification tasks [10, 35]. Specifically, we divide the dataset into a meta-training set, a meta-validation set and a meta-test set. Our aim is to solve an $N$-way $K$-shot video classification problem, where $N$ is the number of classes and $K$ is the number of labeled samples in each class. To train the model, we sample a set of $N$-way $K$-shot tasks from the meta-training set, where a task is also called an episode containing a support set and a query set. The learned model on the meta-training set is evaluated on the meta-test set.

## 4 PROPOSED APPROACH

Fig. 1 illustrates our proposed ensemble learning based architecture, which is composed of (a) a collection of *SlowFast Memory Networks (SFMN)* and (b) a *Choice Controller Network (CCN)*. The SFMN, which is augmented with an external memory unit, aims at predicting diverse classification results at multiple rates, each of which contains a slow pathway and a fast pathway. Next, the CCN makes the final decision by aggregating the results from the ensemble of SFMNs.

### 4.1 SlowFast Memory Networks

We build our base-learner upon the 'SlowFast' strategy which can significantly improve the performance of video classification [8]. Specifically, the SlowFast network contains a slow pathway to capture spatial appearances at a low frame rate, and a fast pathway to extract temporal motion information at a high frame rate. The two pathways are thus parallel and complementary to each other.

As illustrated in Fig. 2, the **slow pathway** is leveraged to capture the representation of a video clip with a large temporal stride $\tau$, indicating that the network processes one out of $\tau$ frames. If the raw video clip has $T \times \tau$ frames, the slow pathway will take $T$ frames as samples. In addition, the **fast pathway** is utilized to extract fine-grained temporal features with a small temporal stride of $\tau/\alpha$, where $\alpha > 1$ is the frame rate ratio. Hence, the fast pathway can sample

$\alpha T$ frames, making it denser than the slow pathway. Furthermore, the fast pathway also utilizes non temporal downsampling layers, and thus captures high-resolution features and maintains sufficient temporal fidelity of the input video clip. Besides, the fast pathway adopts a ratio of $\beta$ ($\beta < 1$) channels to improve the computational efficiency. Subsequently, we fuse the features captured by the two pathways by lateral connections [9, 23] after pool-1, res-2, res-3 and res-4 layers, as shown in Fig. 2. Since the temporal dimensions of the two pathways are different, the lateral connections transform the features from the fast pathway to align with the slow pathway. At last, both pathways' outputs are concatenated after a global average pooling layer, and then fed to the following memory unit to obtain the final result.

To fit in the few-shot video classification task, we introduce an external memory unit inspired by 'memory network' [49] to augment each SlowFast network through a feed-forward network, which is able to store and memorize crucial information of an example video in a relatively long term, even though the given video has been observed only once. Our designed memory unit can provide each SlowFast network with powerful 'memorization' capability by writing new information from videos to memory, and the stored information can be easily accessed and utilized for testing. The SlowFast network retrieves videos' representations from memory using a *read* operation and place videos' features into memory with a *write* operation. Given an input video $x_t$, we produce a corresponding key $k_t$, *i.e.,* a vectorized and normalized hidden representation of $x_t$ concatenated from Slow pathway and Fast pathway, which will be stored in a row of a memory matrix $\Omega_t$. These keys play an crucial role to help our proposed SlowFast memory network capture the optimal video representations during the training process, and hence provide a larger search space to find relatively exact matching in testing. When retrieving a memory, the cosine similarity measurement is adopted to compute the relevance between the key $k_t$ and a particular memory $\Omega_t(i)$ from the $i$-th row of the memory matrix. Formally,

$$\cos(k_t, \Omega_t(i)) = \frac{k_t \cdot \Omega_t(i)}{\|k_t\| \|\Omega_t(i)\|}. \tag{1}$$

Then, we produce a read-weight vector $\sigma_t^r$ corresponding to $k_t$ through a softmax function to guide the memory retrieval as follows:

$$\sigma_t^r(i) \leftarrow \frac{\exp(\cos(k_t, \Omega_t(i)))}{\sum_j \exp(\cos(k_t, \Omega_t(j)))}. \tag{2}$$

Hence, a memory slot $r_t$ corresponding to $k_t$ will be retrieved based on the computed weight vector:

$$r_t \leftarrow \sum_i \sigma_t^r(i)\Omega_t(i). \tag{3}$$

Finally, the retrieved $r_t$ in the memory is utilized as input to a classifier, *i.e.,* a softmax layer, to produce the predicted result for query $x_t$. Thus, our proposed SlowFast memory network can not only capture enough spatial context and exact temporal characteristics in terms of low and high frame rates, but also efficiently record and access the essential representations of examples even when given only once or twice.

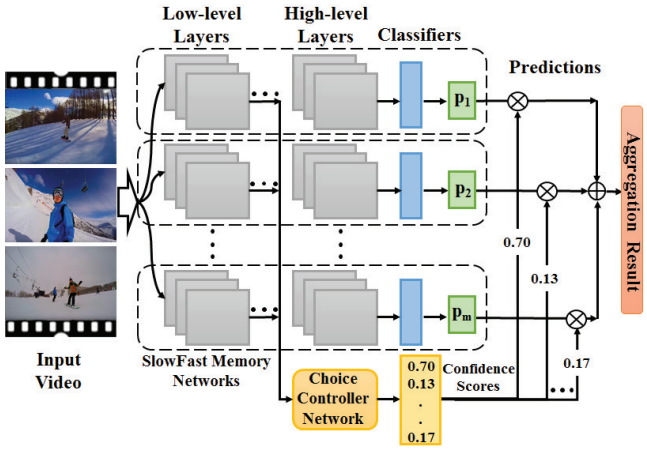


**Figure 3: Illustration of the proposed choice controller network (CCN) for a collection of SlowFast memory networks, where CCN can generate a set of corresponding confidence scores based on the concatenated features of the low-level layers from each network. The best classification result can then be achieved by aggregating the diverse predictions from all the networks.**

## 4.2 Choice Controller Network

Due to scarcity of labeled data in each class with respect to few-shot task, a single SlowFast network would not be fully trained, resulting in a weak classifier. Thus, in order to establish a strong classifier, we propose to learn an ensemble of weak classifiers, which are aggregated by a choice controller network. As illustrated in Fig. 3, the choice controller network is deployed after several layers of the SlowFast network and learns to output the confidence score of each network with a new designed loss function. Hence, the input of our choice controller network is the concatenated features of all the SlowFast memory networks, and the output is an $M$-dimensional vector indicating the confidence scores assigned to the $M$ networks.

More concretely, given $M$ SlowFast memory networks, the parameters of the SlowFast memory networks and the confidence controller network are defined as $\{\theta_m\}_{m=1}^M$ and $\theta_c$, respectively. Meanwhile, the input meta-training dataset is denoted as $\mathcal{D}$ with $N$ categories, and then we sample a mini-batch of video examples referred to as $\mathcal{B} \in \mathcal{D}$, where an input video example is defined as $x_t$. Furthermore, we define the prediction probability of the $m$-th SlowFast network on video $x_t$ as $P_{\theta_m}(\hat{y}_t|x_t)$, and the output confidence score of the confidence controller network is referred to as $[s_t^1, \cdots, s_t^M]$ through a softmax layer to ensure $\sum_{m=1}^M s_t^m = 1$, where $y_t$ and $\hat{y}_t$ are the ground-truth and predicted category for $x_t$, respectively. Formally, the final output probability score of classifying a video example $x_t$ to class $c$ is formulated as:

$$P_{fin}(c|x_t) = \sum_{m=1}^{M} s_t^m P_{\theta_m}(\hat{y}_t = c|x_t). \tag{4}$$

Because of the imbalanced category distribution and limited training data in the dataset, a deep neural network is prone to classify an unseen video to certain categories with high scores

of over confidence. To overcome such problems, we propose a confidence hinge loss in our objective function. Then, the objective of our proposed ensemble learning model can be formulated as follows:

$$\min_{\mu,\theta_m,\theta_c} \mathcal{L}(D) = \sum_{t=1}^{N} \Big[ \sum_{m=1}^{M} \mu_t^m l\big(y_t, P_{\theta_m}(\hat{y}_t|x_t)\big) + l(\mu_t, s_t) + \lambda \sum_{c \neq y_t}^{C} \max\big(P_{fin}(c|x_t) - P_{fin}(y_t|x_t) + \gamma, 0\big) \Big] \quad (5)$$
$$s.t. \sum_{m=1}^{M} \mu_t^m = 1, \forall t; \quad \mu_t^m \in \{0,1\}, \forall t, m,$$

where $\mu_t^m$ refers to the indicator variable, $c$ means the $c$-th category in the dataset, $y_t$ and $\hat{y}_t$ are the ground-truth and predicted result, respectively. Here $\mu_t^m = 1$ denotes that the $m$-th model is the best choice to recognize the $t$-th video example. Meanwhile, $l(\cdot,\cdot)$ denotes the cross entropy function, and $\lambda$ and $\gamma$ are the hyper-parameters that balance the margin-based loss and the confidence margin, respectively.

In Eq. (5), the first term is leveraged to minimize the loss of the most confident or reliable SlowFast model, and make each network can perform better on some particular classes than other base networks according to the indicator variable $\mu_t^m$. The second term enables the choice controller network to make maximally accurate prediction depending on the various outputs of different SlowFast networks, and finally assigns the confidence score to each SlowFast model. The third term is proposed to make the probability of the predicted correct class higher than the incorrect classes, in order to alleviate the overconfidence problem. In other words, if some particular SlowFast networks make wrong predictions with high confidence, the proposed loss is capable of depressing that and promote the SlowFast network to predict the label correctly with a high probability. Furthermore, our proposed objective function would not affect the final prediction even if the maximal probability is too high, which can also preserve the diversity of classifiers in the ensemble.

### 4.3 Training and Inference

**Training:** Our proposed model can be trained in an end-to-end manner by following the commonly-adopted episode training strategy [10, 35]. Algorithm 1 summarizes the whole training procedure of our model based on stochastic gradient descent (SGD). During mini-batch based training, given a query video and the corresponding ground-truth label, each SlowFast memory network retrieves the memory slot from the memory matrix as input to the softmax layer based on Eq. (3). The memory will be cleared by initializing all memory variables with zeros in each episode.

**Inference:** Given a test video example $x$, all the SlowFast memory networks in our ensemble model will produce $M$ diverse classification possibilities, *i.e.,* $P_{\theta_m}(\hat{y}|x)$ $(m = 1, \cdots, M)$, and then the proposed choice controller network will generate $P_{fin}(x)$ by aggregating such outputs to produce the final result. Note that the weights of each SlowFast network are fixed and the content in the memory will be updated with the support set.

**Algorithm 1** Meta-Training Algorithm of Our Ensemble Model

**Input:** Meta-training set $\mathcal{D}$, and trade-off hyper-parameters: $\lambda$, $\gamma$;
**Output:** The trained ensemble model;

Randomly initialize parameters of each SlowFast memory network and choice controller network, *i.e.,* $\{\theta_m\}_{m=1}^M$ and $\theta_c$
**repeat**
  Sample a N-way K-shot task batch $\mathcal{B} \in \mathcal{D}$
  **for** $m = 1 \rightarrow M$ **do**
    // Output of each SlowFast memory network in a batch
    $P_{\theta_m}(\mathcal{B}) \rightarrow \hat{y}_{m,1}, \cdots, \hat{y}_{m,|\mathcal{B}|}$;
  **end for**
  **for** $t = 1 \rightarrow \mathcal{B}$ **do**
    $\mu_t^m = 0$, $m = 1, \cdots, M$;
    // Choose the lowest error model for each example
    $m^* \leftarrow \arg\min_{m \in [1,\cdots,M]} l(y_t, \hat{y}_{m,t})$, and set $\mu_t^{m^*} = 1$;
    // Determine the confidence scores from choice controller network based on Eq. (5)
    $s_t = [s_t^1, \cdots, s_t^M]$
    // Calculate the best prediction
    $P_{fin}(\hat{y}_t|x_t) = \sum_{m=1}^M s_t^m P_{\theta_m}(\hat{y}_t|x_t)$
    // Calculate the gradient w.r.t. parameters $\theta_c$ by Backprop optimization
    $\partial \mathcal{L}(x_t)/\partial \theta_c$
    // Calculate the gradient w.r.t. parameters $\{\theta_m\}_{m=1}^M$ by Backprop optimization
    $\partial \mathcal{L}(x_t)/\partial \theta_m$
  **end for**
  Update the parameters in the whole model
**until** converage

## 5 EXPERIMENTAL RESULTS

In this section, we conduct extensive experiments to evaluate our proposed ensemble learning network for few-shot video classification tasks on the widely-adopted **Kinetics** [19] and **Charades** [40] datasets. We conduct comprehensive comparison with previous methods and extensive ablation study to gain insight into the effectiveness of our model.

### 5.1 Experimental Settings

**Datasets:** The **Kinetics** [19] dataset contains 306,245 videos from 400 categories. All video examples on this dataset belong to various kinds of human actions. We adopt the same experimental setting as in [56], and randomly choose 100 classes, each of which includes 100 video examples. The **Charades** dataset [40] includes around 9.8k videos as the training set, and 1.8k videos as the validation set from 157 classes, each of which lasts more than 30 seconds on average. Similar to the settings on Kinetics, we randomly choose 100 classes and 100 examples per category from Charades. On both datasets, we split all the categories into 64, 12 and 24 classes as the meta-training, meta-validation and meta-test sets in our experiments, respectively.

**Evaluation Metrics:** In few-shot learning, the evaluation is performed in terms of $N$-way $K$-shot classification tasks in the meta-test set, from which we randomly sample a set of $N$-way $K$-shot tasks, and an unlabeled query example belonging to one of such $N$

**Table 1: Performance comparison and ablation study of our full model (M=10), baseline models, and the state-of-the-art methods on Kinetics and Charades in terms of 5-way few-shot video classification on the meta-test set.**

| Methods | Kinetics [19] | | | | | Charades [40] | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1-shot | 2-shot | 3-shot | 4-shot | 5-shot | 1-shot | 2-shot | 3-shot | 4-shot | 5-shot |
| ResNet-50 RGB | 28.7 | 36.8 | 42.6 | 46.2 | 48.6 | 13.2 | 21.7 | 26.5 | 29.8 | 30.3 |
| ResNet-50 Flow | 24.4 | 27.3 | 29.8 | 32.0 | 33.1 | 10.5 | 12.1 | 13.9 | 15.7 | 17.2 |
| LSTM RGB | 28.9 | 37.5 | 43.3 | 47.1 | 49.0 | 13.9 | 22.5 | 27.6 | 30.7 | 31.1 |
| Nearest-Finetune | 48.2 | 55.5 | 59.1 | 61.0 | 62.6 | 15.7 | 23.7 | 29.1 | 31.9 | 33.2 |
| Nearest-Pretrain | 51.1 | 60.4 | 64.8 | 67.1 | 68.9 | 16.2 | 24.0 | 29.8 | 32.5 | 33.9 |
| MatchingNet [46] | 53.3 | 64.3 | 69.2 | 71.8 | 74.6 | 18.9 | 25.2 | 31.7 | 33.2 | 35.1 |
| MAML [10] | 54.2 | 65.5 | 70.0 | 72.1 | 75.3 | 19.2 | 26.8 | 32.3 | 34.5 | 36.7 |
| LSTM Embed | 57.6 | 67.9 | 72.8 | 74.8 | 76.2 | 21.8 | 29.1 | 34.3 | 36.1 | 38.2 |
| Plain CMN [18] | 57.3 | 67.5 | 72.5 | 74.7 | 76.0 | 21.3 | 28.7 | 33.8 | 35.6 | 37.9 |
| Video CMN [56] | 60.5 | 70.0 | 75.6 | 77.3 | 78.9 | 23.6 | 31.9 | 36.2 | 37.8 | 40.6 |
| **Our Full Model** | **63.7** | **73.9** | **79.5** | **80.9** | **83.1** | **25.7** | **33.5** | **38.7** | **39.8** | **42.3** |
| Ensemble-Max | 61.2 | 71.3 | 76.9 | 78.5 | 80.2 | 25.1 | 32.9 | 38.1 | 39.2 | 41.6 |
| Ensemble-Avg | 61.9 | 71.8 | 77.5 | 79.1 | 80.5 | 25.3 | 33.1 | 38.3 | 39.5 | 41.9 |
| Ours w/o Ensemble Learning | 58.0 | 68.1 | 74.2 | 75.3 | 78.2 | 22.8 | 29.3 | 35.7 | 37.6 | 39.2 |
| Ours w/o SlowFast Memory Network | 59.2 | 69.5 | 74.9 | 76.7 | 79.8 | 23.2 | 30.5 | 36.2 | 38.1 | 40.7 |
| Ours w/o Memory Unit | 60.7 | 70.6 | 76.2 | 77.9 | 79.5 | 24.3 | 32.2 | 37.5 | 38.3 | 41.2 |

categories needs to be classified during test. Similar to [10, 46, 56], we also adopt the mean accuracy by randomly sampling 20,000 episodes across all our experiments, and report the performance of 1-shot, 2-shot, 3-shot, 4-shot, and 5-shot in terms of 5-way classification tasks.

**Compared Methods:** We compare our proposed framework with several state-of-the-art approaches, *i.e.,* **MatchingNet** [46], **MAML** [10], **Plain CMN** [18], and **Video CMN** [56]. Especially, Video CMN [56] is the only one mainly focusing on few-shot video classification, while the others are proposed to address few-shot image classification but trained on the video datasets in our experiments. Additionally, we introduce several baseline models including ResNet-50 with RGB frames and stacked optical flow images as input, respectively ('**ResNet-50 RGB**'/'**ResNet-50 Flow**'), LSTM with RGB frames as input ('**LSTM RGB**'), a baseline model utilized nearest neighbour with fine-tuned and pre-trained, respectively ('**Nearest-Finetune**'/'**Nearest-Pretrain**'). Another baseline '**LSTM Embed**' is a variant of Video CMN by replacing the embedding module in Video CMN with an LSTM unit.

## 5.2 Implementation Details

Our proposed model is implemented with the PyTorch library, and run on eight Nvidia GeForce GTX 1080Ti GPUs. In the training process, we randomly sample data from the meta-train set to make the experimental results more solid. Following the episode training mechanism, there are totally 20,000 episodes, each of which includes a support set and a query set. We adopt SGD to train our model and set the batch size to 32. The initial learning rate is set to $1\times10^{-4}$, and then reduced by half for every 5,000 episodes. During test, 200 episodes are randomly selected from the test set to obtain the top-1 mean accuracy. Note that our proposed whole model including ensemble SlowFast memory networks and choice controller network is trained in an end-to-end manner from scratch without fine-tuning process. In all experiments, we leverage ResNet-50 [15] as our backbone model of each SlowFast memory network to capture the frame-level features of a video, and conduct experiments on M networks in the ensemble ($M = \{1, 2, 3, 5, 10, 20\}$). We implement the choice controller network with 3 fully-connected layers, and deploy it after the last convolutional layer of each SlowFast network. For preprocessing, each frame of input video is randomly cropped to 224×224 pixels. Besides, we randomly sample a clip of $\alpha T \times \tau$ frames from the video example, and input $T$ and $\alpha T$ frames to the network. As for the hyper-parameters in our proposed model, we set $\tau = 16$, $\alpha = 8$ and $\beta = 1/8$ in all SlowFast memory networks. And we set $\lambda = 1$, $\gamma = 0.3/0.6$ on the Kinetics/Charades datasets, respectively, in the choice controller network. These hyper-parameters are tuned on the meta-validation set, and the training process would stop once the accuracy starts decreasing.

## 5.3 Results and Analysis

Table 1 summarizes the performance comparison of our proposed approach (10 networks in the ensemble), other baselines, and current state-of-the-art methods w.r.t. 1-shot, 2-shot, 3-shot, 4-shot, and 5-shot in terms of 5-way classification tasks on the Kinetics and Charades datasets. From the table, we can observe that our proposed model achieves the best performance across all few-shot classification tasks on both datasets, which verifies the significance and effectiveness of the ensemble learning strategy and the memory-augmented SlowFast networks. We find that fine-tuned ResNet-50 with RGB frames and optical flow images on the meta-training

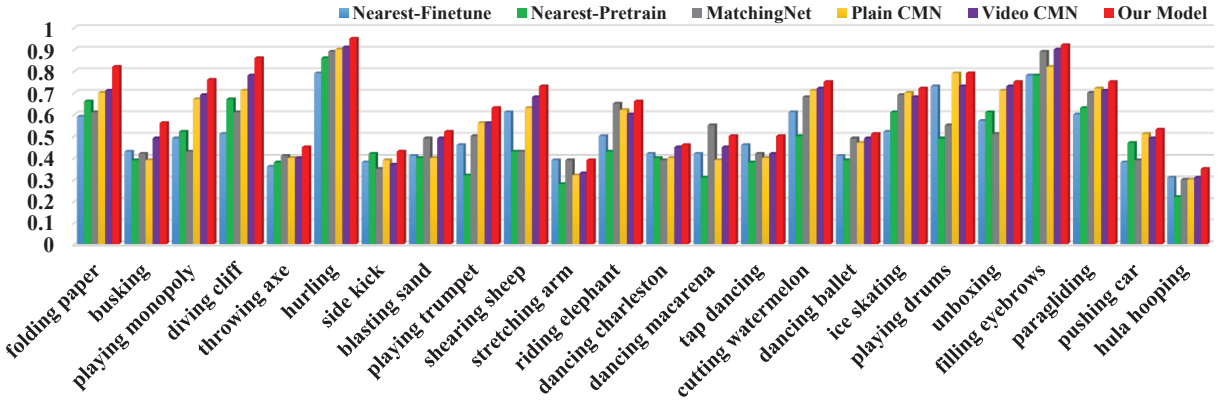


**Figure 4: Comparison of per-class recognition rate of our ensemble method (M=10) with the state-of-the-art approaches on Kinetics in terms of 5-way 1-shot video classification.**

dataset cannot help to improve the result of few-shot video classification. This could be attributed to the overfitting of the fine-tuned model on the meta-training set, which has very poor generalization capability to the novel classes of the meta-test set. Compared with these baselines, our proposed model enhances the performance with a very large margin, again demonstrating the effectiveness of our ensemble learning based model. Furthermore, our proposed method outperforms the traditional few-shot learning methods (*e.g.,* MatchingNet, MAML and Plain CMN) by more than ~10% across all shots on both datasets. This is due to our proposed SlowFast memory networks, which have a stronger capability to capture multi-rate discriminative representations of videos, while other state-of-the-art methods are only dedicated to the image domain. In addition, our proposed approach also outperforms the other video-based few-shot learning method 'Video CMN' by about 3% and 2% in terms of all shots on Kinetics and Charades, respectively.

Furthermore, we report the top-1 mean accuracy of each class w.r.t. 5-way 1-shot video classification in Fig. 4. We can clearly see that our ensemble model with 10 networks outperforms other state-of-the-art approaches in most of the video categories. For instance, our model obtains more than 80% accuracy with respect to 'folding paper', 'diving cliff', 'hurling' and 'filling eyebrows'. It could be attributed to that our ensemble learning based model is able to capture the relevance between base-learners and leverage distribution discrepancies for video classification, and thus our proposed choice controller network can make the best prediction.

## 5.4 Ablation Study

We perform further ablation studies to verify the effectiveness of different components in our framework.

**Importance of choice controller network.** In our proposed model, the introduced choice controller network is one of the main contributions. It can generate various confidence scores for each SlowFast memory network and adaptively aggregate all results to achieve the best choice. Here, we propose two alternative operations in ensemble learning, *i.e.,* only adopting the maximum or averaged confidence result of particular SlowFast memory network as the final prediction, denoted as 'Ensemble-Max' and 'Ensemble-Avg' in Table 1 and Fig. 5, respectively. We can clearly observe from Table 1

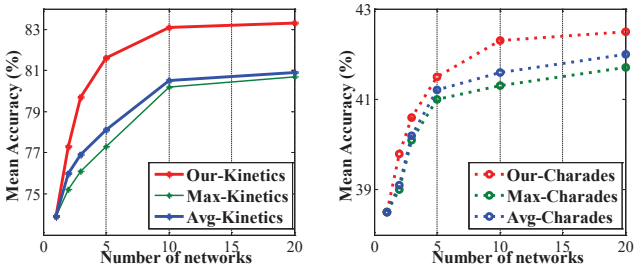


**Figure 5: 5-shots performance of different ensemble strategies (*i.e.,* 'Our', 'Max' and 'Avg') for various numbers of SlowFast memory networks (M=1, 2, 3, 5, 10, 20) in our proposed approach on Kinetics and Charades.**

and Fig. 5 that our proposed full model with choice controller network outperforms both the 'Ensemble-Max' and 'Ensemble-Avg' across all shots on both datasets no matter employing how many networks in the ensemble, especially more than nearly 2% on Kinetics. This indicates that aggregating the final result based on various confidence scores is better than directly selecting the maximum or averaged result in an ensemble, because our proposed strategy can greatly preserve the diversity of each base learner in an ensemble. Furthermore, our full model can beat a single SlowFast memory network (denoted as 'Ours w/o Ensemble Learning' in Table 1) by a large margin, demonstrating the effectiveness of the proposed ensemble learning architecture again. It is worth noting that our model without ensemble learning ('Ours w/o Ensemble Learning') leads to the worst performance across two datasets compared against our model without SlowFast memory network ('Ours w/o SlowFast Memory Network') as shown in Table 1, showing that our designed ensemble learning structure plays a more important role in improving the performance of the whole model. Additionally, during each test case, we calculate the possibility residuals between the final probability aggregated by choice controller network and predicted probability of each SlowFast memory network (define as $R_m(\hat{y}|x) = P_{fin}(\hat{y}|x) - P_{\theta_m}(\hat{y}|x)$). The probability residuals near zero indicate that the m-th model is assigned a high confidence score on these samples, while the probability residuals near 1 denote the m-th model has low confidence for those data. The

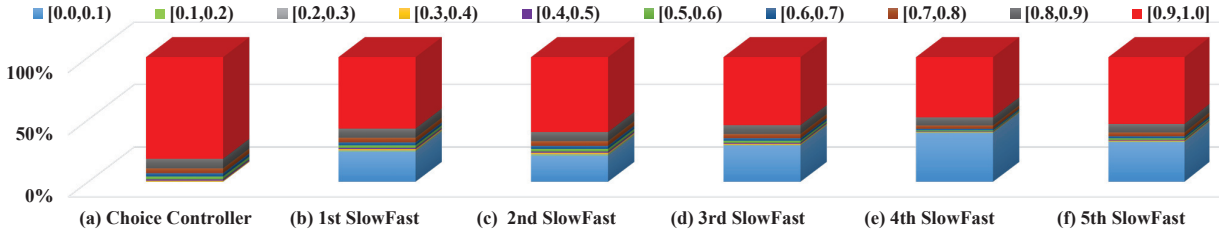


**Figure 6: The prediction probability distribution of our ensemble model (5 networks) tested on Kinetic meta-test set. (a) The $P_{fin}$ of choice controller network which is achieved by aggregating all SlowFast memory models. (b)-(f) show the probability residual between $P_{fin}$ and $P_m$ for m=1, 2, 3, 4 and 5 respectively. The y axis denotes the percentage of prediction probability distribution across all test cases.**

**Table 2: Performance of different way and shot settings of few-shot video classification on Kinetics.**

| | 1-shot | 2-shot | 3-shot | 4-shot | 5-shot |
|---|---|---|---|---|---|
| 5-way | **63.7** | **73.9** | **79.5** | **80.9** | **83.1** |
| 6-way | 59.6 | 68.2 | 75.1 | 77.5 | 80.9 |
| 7-way | 57.3 | 66.8 | 73.2 | 75.0 | 77.8 |
| 8-way | 53.9 | 63.1 | 70.5 | 72.6 | 75.1 |

**Table 3: Performance of various memory sizes in our proposed SlowFast Memory Network on Kinetics.**

| | 1-shot | 2-shot | 3-shot | 4-shot | 5-shot |
|---|---|---|---|---|---|
| Mem-32 | 56.2 | 65.6 | 71.8 | 73.0 | 75.5 |
| Mem-64 | 59.7 | 68.5 | 75.2 | 76.3 | 79.9 |
| Mem-128 | 61.2 | 71.5 | 77.3 | 78.6 | 81.5 |
| Mem-512 | **63.7** | **73.9** | **79.5** | **80.9** | **83.1** |
| Mem-2048 | 63.5 | 73.6 | 79.2 | 80.5 | 82.8 |

distribution of possibility residual results on Kinetics meta-test set are shown in Fig. 6. We can find from Fig. 6 (a) that most of final prediction probability lies in $[0.9, 1.0]$, denoting that our proposed choice controller network is able to make the best decision with high confidence. From Fig. 6 (b)-(f), we can see that each SlowFast memory network can only be assigned high confidence to make the final decision on limit samples and categories, demonstrating that our proposed ensemble architecture can effectively exploit the diversity and improve the final prediction with the help of choice controller network.

**Number of models in ensemble learning.** We also examine the effect of the number of SlowFast memory networks on the overall performance. The result is shown in Fig. 5, where we can see that the performance increases along with the increasing number of networks in the ensemble, but the increase rate would decrease when the number exceeds 10. This is because it becomes difficult to converge with so many networks, and a larger number of networks in the ensemble will definitely lead to more time-consuming computations. Hence, we generally choose 10 SlowFast memory networks in the ensemble considering a good trade-off between classification accuracy and computational efficiency.

**Effectiveness of SlowFast memory network.** As denoted in Table 1, our full model with SlowFast memory network outperforms our model without SlowFast memory network (denoted as 'Ours w/o SlowFast Memory Network' in Table 1, which means our model only utilizes ResNet-50 as the backbone model) across all shot tasks, indicating that the proposed SlowFast memory network can effectively capture better spatial-temporal information through multi-rate processing than ResNet. Meanwhile, the full model outperforms our model without memory unit (denoted as 'Ours w/o Memory Unit' in Table 1) on both datasets. It should be attributed to the proposed memory unit which can significantly improve the few-shot classification performance by efficiently recording and retrieving the information from a generated memory matrix.

**N-way few-shot classification.** In addition to 5-way few-shot video classification, we examine the performance of N-way classification. As shown in Table 2, the performance degrades along with the increasing of N, demonstrating it remains a challenge for few-shot learning with too many ways. In addition, the performances of N-way one-shot tasks still have a large space to improve further.

**Memory size.** We report the results of different memory sizes in the memory unit in Table 3 and observe that our model augmented with the memory unit can obtain the best performance when the memory size is set to 512. If the memory size is too small, it is not enough to store the information of novel category data; if it is too large, it leads to too much noise and thus the performance cannot be further improved either.

Finally, we note that our proposed ensemble approach achieves better performance at the expense of training multiple learners. Thus, this is a reasonable trade-off.

## 6 CONCLUSION

In this paper, we present an ensemble learning framework with multi-rate SlowFast memory networks for few-shot video classification. We emphasize and verify the importance of learning spatial-temporal information of a video at different frame rates and learning various confidence scores to exploit the diversity of multiple base-learners. Both of them are beneficial for few-shot video classification and largely boost the performance. Extensive experiments on two popular benchmarks show competitive results compared with several state-of-the-art models and baselines. In the future, how to apply the proposed model to few-shot cross-modal video retrieval and video captioning deserves further exploration.

## REFERENCES


[1] Leo Breiman. 1996. Bagging predictors. *Machine learning* 24, 2 (1996), 123–140.
[2] Spencer Cappallo and Cees GM Snoek. 2017. Future-Supervised Retrieval of Unseen Queries for Live Video. In *Proceedings of the 25th ACM international conference on Multimedia (MM)*. ACM, 28–36.
[3] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. CVPR*. IEEE, 6299–6308.
[4] Xiongtao Chen, Wenmin Wang, Jinzhuo Wang, and Weimian Li. 2017. Learning object-centric transformation for video prediction. In *Proceedings of the 25th ACM international conference on Multimedia (MM)*. ACM, 1503–1512.
[5] Donghyeon Cho, Yunjae Jung, Francois Rameau, Dahun Kim, Sanghyun Woo, and In So Kweon. 2019. Video Retargeting: Trade-off between Content Preservation and Spatio-temporal Consistency. In *Proceedings of the 27th ACM International Conference on Multimedia (MM)*. ACM, 882–889.
[6] Navneet Dalal, Bill Triggs, and Cordelia Schmid. 2006. Human detection using oriented histograms of flow and appearance. In *Proc. ECCV*. Springer, 428–441.
[7] Shancheng Fang, Hongtao Xie, Zheng-Jun Zha, Nannan Sun, Jianlong Tan, and Yongdong Zhang. 2018. Attention and language ensemble for scene text recognition with convolutional sequence modeling. In *Proceedings of the 26th ACM international conference on Multimedia (MM)*. ACM, 248–256.
[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proc. ICCV*. IEEE, 6202–6211.
[9] Christoph Feichtenhofer, Axel Pinz, and Richard Wildes. 2016. Spatiotemporal residual networks for video action recognition. In *Proc. NIPS*. 3468–3476.
[10] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. ICML*.
[11] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning*. Vol. 1. Springer series in statistics New York.
[12] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics* 28, 2 (2000), 337–407.
[13] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. 2018. Watch, think and attend: End-to-end video classification via dynamic knowledge evolution modeling. In *Proceedings of the 26th ACM international conference on Multimedia (MM)*. ACM, 690–699.
[14] Victor Garcia and Joan Bruna. 2018. Few-shot learning with graph neural networks. *Proc. ICLR* (2018).
[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *Proc. ECCV*. Springer, 630–645.
[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[17] Linxi Jiang, Xingjun Ma, Shaoxiang Chen, James Bailey, and Yu-Gang Jiang. 2019. Black-box adversarial attacks on video recognition models. In *Proceedings of the 27th ACM International Conference on Multimedia (MM)*. ACM, 864–872.
[18] Łukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. 2017. Learning to remember rare events. In *Proc. ICLR*.
[19] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).
[20] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld. 2008. Learning realistic human actions from movies. In *Proc. CVPR*. IEEE, 1–8.
[21] Dong Li, Ting Yao, Zhaofan Qiu, Houqiang Li, and Tao Mei. 2019. Long Short-Term Relation Networks for Video Action Detection. In *Proceedings of the 27th ACM International Conference on Multimedia (MM)*. ACM, 629–637.
[22] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. 2017. Attention transfer from web images for video recognition. In *Proceedings of the 25th ACM international conference on Multimedia (MM)*. ACM, 1–9.
[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proc. CVPR*. IEEE, 2117–2125.
[24] Jakub Lokoč, Gregor Kovalčik, Tomáš Souček, Jaroslav Moravec, and Přemysl Čech. 2019. A framework for effective known-item search in video. In *Proceedings of the 27th ACM International Conference on Multimedia (MM)*. ACM, 1777–1785.
[25] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. 2018. A simple neural attentive meta-learner. *Proc. ICLR* (2018).
[26] Nikunj Chandrakant Oza and Stuart Russell. 2001. *Online ensemble learning*. University of California, Berkeley.
[27] Mengshi Qi, Weijian Li, Zhengyuan Yang, Yunhong Wang, and Jiebo Luo. 2019. Attentive Relational Networks for Mapping Images to Scene Graphs. In *Proc. CVRR*. IEEE, 3957–3966.
[28] Mengshi Qi, Jie Qin, Annan Li, Yunhong Wang, Jiebo Luo, and Luc Van Gool. 2018. stagnet: An attentive semantic RNN for group activity recognition. In *Proc. ECCV*. Springer, 101–117.
[29] Mengshi Qi, Jie Qin, Yu Wu, and Yi Yang. 2020. Imitative Non-Autoregressive Modeling for Trajectory Forecasting and Imputation. In *Proc. CVRR*. IEEE, 12736–12745.
[30] Mengshi Qi, Yunhong Wang, and Annan Li. 2017. Online cross-modal scene retrieval by binary representation and semantic graph. In *Proceedings of the 25th ACM international conference on Multimedia (MM)*. ACM, 744–752.
[31] Mengshi Qi, Yunhong Wang, Annan Li, and Jiebo Luo. 2018. Sports Video Captioning by Attentive Motion Representation based Hierarchical Recurrent Neural Networks. In *Proceedings of the 1st International Workshop on Multimedia Content Analysis in Sports (MMSports)*. ACM, 77–85.
[32] Mengshi Qi, Yunhong Wang, Annan Li, and Jiebo Luo. 2020. STC-GAN: Spatio-Temporally Coupled Generative Adversarial Networks for Predictive Scene Parsing. *IEEE Transactions on Image Processing* 29 (2020), 5420–5430.
[33] Mengshi Qi, Yunhong Wang, Jie Qin, and Annan Li. 2019. KE-GAN: Knowledge Embedded Generative Adversarial Networks for Semi-Supervised Scene Parsing. In *Proc. CVPR*. IEEE, 5237–5246.
[34] Xufeng Qian, Yueting Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, and Jun Xiao. 2019. Video relation detection with spatio-temporal graph. In *Proceedings of the 27th ACM International Conference on Multimedia (MM)*. ACM, 84–93.
[35] Sachin Ravi and Hugo Larochelle. 2017. Optimization as a model for few-shot learning. In *Proc. ICLR*.
[36] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. Meta-learning with memory-augmented neural networks. In *Proc. ICML*. 1842–1850.
[37] Robert E Schapire. 2003. The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*. Springer, 149–171.
[38] Robert E Schapire and Yoram Singer. 1999. Improved boosting algorithms using confidence-rated predictions. *Machine learning* 37, 3 (1999), 297–336.
[39] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. 2017. Video visual relation detection. In *Proceedings of the 25th ACM international conference on Multimedia (MM)*. ACM, 1300–1308.
[40] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proc. ECCV*. Springer, 510–526.
[41] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Proc. NIPS*. 568–576.
[42] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Proc. NIPS*. 4077–4087.
[43] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *Proc. CVPR*. IEEE, 1199–1208.
[44] Kai Tian, Yi Xu, Shuigeng Zhou, and Jihong Guan. 2019. Versatile multiple choice learning and its application to vision computing. In *Proc. CVPR*. IEEE, 6349–6357.
[45] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proc. ICCV*. IEEE, 4489–4497.
[46] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. In *Proc. NIPS*. 3630–3638.
[47] Heng Wang and Cordelia Schmid. 2013. Action recognition with improved trajectories. In *Proc. ICCV*. IEEE, 3551–3558.
[48] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *Proc. ECCV*. Springer, 20–36.
[49] Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916* (2014).
[50] Haoze Wu, Zheng-Jun Zha, Xin Wen, Zhenzhong Chen, Dong Liu, and Xuejin Chen. 2019. Cross-Fiber Spatial-Temporal Co-enhanced Networks for Video Action Recognition. In *Proceedings of the 27th ACM International Conference on Multimedia (MM)*. ACM, 620–628.
[51] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. 2015. Beyond short snippets: Deep networks for video classification. In *Proc. CVPR*. IEEE, 4694–4702.
[52] Chenrui Zhang, Xiaoqing Lyu, and Zhi Tang. 2019. TGG: Transferable Graph Generation for Zero-shot and Few-shot Learning. In *Proceedings of the 27th ACM International Conference on Multimedia (MM)*. ACM, 1641–1649.
[53] Zhaoyang Zhang, Zhanghui Kuang, Ping Luo, Litong Feng, and Wei Zhang. 2018. Temporal sequence distillation: Towards few-frame action recognition in videos. In *Proceedings of the 26th ACM international conference on Multimedia (MM)*. ACM, 257–264.
[54] Zheng Zhang, Zhihui Lai, Zi Huang, Wai Keung Wong, Guo-Sen Xie, Li Liu, and Ling Shao. 2019. Scalable supervised asymmetric hashing with semantic and latent factor embedding. *IEEE Transactions on Image Processing* 28, 10 (2019), 4803–4818.
[55] Zheng Zhang, Li Liu, Fumin Shen, Heng Tao Shen, and Ling Shao. 2018. Binary multi-view clustering. *IEEE transactions on pattern analysis and machine intelligence* 41, 7 (2018), 1774–1782.
[56] Linchao Zhu and Yi Yang. 2018. Compound memory networks for few-shot video classification. In *Proc. ECCV*. Springer, 751–766.
[57] Yaochen Zhu, Zhenzhong Chen, and Feng Wu. 2019. Multimodal Deep Denoise Framework for Affective Video Content Analysis. In *Proceedings of the 27th ACM International Conference on Multimedia (MM)*. ACM, 130–138.