Active Exploration of Multimodal Complementarity for Few-Shot Action Recognition

Yuyang Wanyan[1,2], Xiaoshan Yang[1,2,3], Chaofan Chen[4], Changsheng Xu[1,2,3*]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences (CASIA)
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS) [3]Peng Cheng Laboratory, China
[4]School of Information Science and Technology, University of Science and Technology of China (USTC)
wanyanyuyang2021@ia.ac.cn, xiaoshan.yang@nlpr.ia.ac.cn, chencfbupt@gmail.com, csxu@nlpr.ia.ac.cn

Abstract

Recently, few-shot action recognition receives increasing attention and achieves remarkable progress. However, previous methods mainly rely on limited unimodal data (e.g., RGB frames) while the multimodal information remains relatively underexplored. In this paper, we propose a novel Active Multimodal Few-shot Action Recognition (AMFAR) framework, which can actively find the reliable modality for each sample based on task-dependent context information to improve few-shot reasoning procedure. In meta-training, we design an Active Sample Selection (ASS) module to organize query samples with large differences in the reliability of modalities into different groups based on modality-specific posterior distributions. In addition, we design an Active Mutual Distillation (AMD) to capture discriminative task-specific knowledge from the reliable modality to improve the representation learning of unreliable modality by bidirectional knowledge distillation. In meta-test, we adopt Adaptive Multimodal Inference (AMI) to adaptively fuse the modality-specific posterior distributions with a larger weight on the reliable modality. Extensive experimental results on four public benchmarks demonstrate that our model achieves significant improvements over existing unimodal and multimodal methods.

1. Introduction

Over the past years, action recognition [20, 34, 52, 73] has achieved significant progress with the emerge of deep learning. However, these existing deep methods require a large amount of labeled videos to guarantee their performance. In practice, it is sometimes expensive or even impossible to collect abundant annotated data, which limits the effectiveness of supervised methods. In order to deal with this problem, more and more researchers begin to focus on the few-shot action recognition (FSAR) task, which aims at classifying unlabeled videos (query set) from novel action classes with the help of only a few annotated samples (support set).

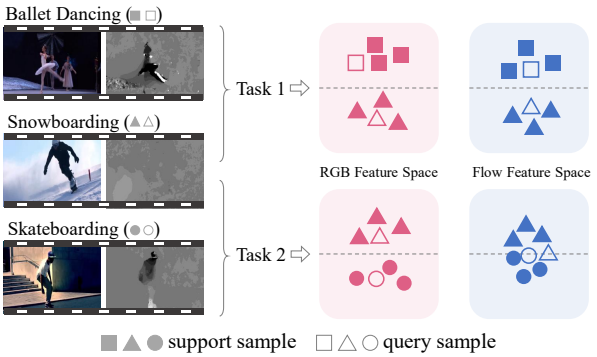

Figure 1. Illustration of multimodal few-shot action recognition task. The main challenge is that the contribution of a specific modality highly depends on task-specific contextual information.

Recently, researchers have proposed many promising few-shot action recognition methods, which can be roughly divided into two groups: data augmentation-based methods and alignment-based methods. Data augmentation-based methods try to generate additional training data [18], self-supervision signals [72] or auxiliary information [22, 69] to promote robust representation learning. Alignment-based methods [5, 8, 44, 58, 66, 69, 72] focus on matching the video frames or segments in the temporal or spatial dimension to measure the distance between query and support samples in a fine-grained manner.

Although existing few-shot action recognition methods have achieved remarkable performance, they mainly rely on limited unimodal data (e.g. RGB frames) that are always insufficient to reflect complex characteristics of human actions. When learning novel concepts from a few samples, humans have the ability to integrate the multimodal perceptions (e.g. appearance, audio and motion) to enhance the recognition procedure. In addition, in conventional action recognition, many top-performing methods [23, 46, 56, 60]

*corresponding author: Changsheng Xu.

always involve multiple modalities (e.g. vision, optical flow and audio) which can provide complementary information to comprehensively identify different actions. Whereas, the multimodal information remains relatively underexplored in few-shot action recognition where the data scarcity issue magnifies the defect of unimodal data.

In this paper, we study multimodal few-shot action recognition task, where the query and support samples are multimodal as shown in Figure 1. With multimodal data, we can alleviate the data scarcity issue through the complementarity of different modalities. However, exploring the multimodal complementarity in few-shot action recognition is nontrivial. On the one hand, although there are many widely used methods for fusing multimodal data, e.g., early fusion [47], late fusion [38, 64], it is still questionable whether existing methods are suitable to be directly applied in the few-shot scenario where only a few samples are available for each action class. On the other hand, the contribution of a specific modality is not consistent for different query samples and it highly depends on the contextual information of both query and support samples in each few-shot task. For example, as shown in Figure 1, if the few-shot task is to identify query samples from the two action classes of `Snowboarding` and `Ballet dancing`, the RGB data and optical flow are equally important and they can complement each other well to distinguish these two classes. In contrast, for the two action classes of `Snowboarding` and `Skateboarding`, the optical flow cannot provide useful discriminative features to complement the vision information or even harm the few-shot recognition performance due to the motion resemblance between these two actions. Therefore, we argue that it requires a task-dependent strategy for exploring the complementarity between different modalities in few-shot action recognition.

In order to reasonably take advantage of the complementarity between different modalities, we propose an **Active Multimodal Few-shot Action Recognition (AMFAR)** framework inspired by active learning [6], which can actively find the more reliable modality for each query sample to improve the few-shot reasoning procedure. AMFAR adopts the episode-wise learning framework [53, 63], where each episode has a few labeled support samples and the unlabeled query samples that need to be recognized. **In each episode of the meta-training**, we firstly adopt modality-specific backbone networks to extract the multimodal representations for query samples and the prototypes of different actions for support samples. We further compute the modality-specific posterior distributions based on query-to-prototype distances. Then, we adopt **Active Sample Selection (ASS)** to organize query samples with large differences in the reliability of two modalities into two groups, i.e., RGB-dominant group that contains samples where the RGB modality is more reliable for conducting action recognition in the current episode, and Flow-dominant group where the optical flow is more reliable. For each query sample, the reliability of a specific modality is estimated according to certainties of the modality-specific posterior distribution. Next, we design an **Active Mutual Distillation (AMD)** mechanism to capture discriminative task-specific knowledge from the reliable modality to improve the representation learning of unreliable modality by bidirectional knowledge guiding streams between modalities. For each query in the RGB-dominant group, the RGB modality is regarded as teacher while the optical flow is regarded as student, and the query-to-prototype relation knowledge is transferred from the teacher to the student with a distillation constraint. Simultaneously, for each query in the Flow-dominant group, optical flow is regarded as teacher while RGB is regarded as student, and the knowledge distillation is conducted in the opposite direction. **In the meta-test phase**, we adopt **Adaptive Multimodal Inference (AMI)** to conduct the few-shot inference for each query sample by adaptively fusing the posterior distributions predicted from different modalities with a larger weight on the reliable modality.

In summary, the main contributions of this paper are fourfold: 1) We exploit the natural complementarity between different modalities to enhance the few-shot action recognition procedure by actively finding the more reliable modality for each query sample. To our best knowledge, we are the first to adopt the idea of active learning to explore the multimodal complementarity in few-shot learning. 2) We propose an active mutual distillation strategy to transfer task-dependent knowledge learned from the reliable modality to guide the representation learning for the unreliable modality, which can improve the discriminative ability of the unreliable modality with the help of the multimodal complementarity. 3) We propose an adaptive multimodal few-shot inference approach to fuse modality-specific results by paying more attention to the reliable modality. 4) We conduct extensive experiments on four challenging datasets and the results demonstrate that the proposed method outperforms existing unimodal and multimodal methods.

2. Related Work

Few-shot Learning. Few-shot learning aims to recognize unseen concepts with only a few labeled training samples. The majority of few-shot learning methods can be divided into two main groups: optimization-based [1, 21, 36, 55] and metric-based methods [11, 12, 51, 53, 57, 63, 68, 71]. Optimization-based methods (e.g., MAML [21]) learn an optimizer to adapt to new tasks with limited training samples. Metric-based methods (e.g., Prototypical [53] and Matching [63] Networks) learn a common metric space for both seen and novel classes and compare query and support samples through a distance in the learned metric space.

In recent years, few-shot learning has achieved great success in many conventional tasks, such as image classification [3, 16], object detection [19, 30, 31], and segmentation [35, 65]. However, most existing methods focus on unimodal data, while only a few works consider multimodal data [17, 39–41, 43, 61]. For example, several methods [39, 40] enrich the low populated visual embedding by leveraging auxiliary text data during training to deal with few-shot image classification at test time. Dong et al. [17] focus on modeling the relationship between images and texts to solve few-shot image captioning and visual question answering. Tsimpoukelli et al. [61] transfer the few-shot learning ability of pretrained language models to downstream tasks (e.g. VQA).

Few-shot Action Recognition. Most existing few-shot action recognition methods [5, 8, 22, 33, 44, 58, 66, 72, 74] adopt the metric learning strategy to help estimate the distances between the query and support samples in a unified metric/feature space for few-shot inference. For example, OTAM [8] aligns the query-support pair with a DTW algorithm that exploits long-term temporal ordering. TRX [44] and STRM [58] model temporal relations by representing the video as tuples consisting of a few sparse frames and compare the similarity of the query-support pair in a part-based manner. Inspired by data augmentation, there are also methods to learn distinguishable action-specific characteristics with extra self-supervisory information or generated training data. For example, ARN [72] introduces spatial and temporal self-supervision to learn a robust video representation. Other methods [18, 70] leverage generative adversarial network (GAN) to synthesize additional examples for novel categories. Perhaps the most related works to our paper are [22, 69] that generate auxiliary information to promote the visual representation of image sequences in videos. Specifically, AMeFu-Net [22] introduces depth as a carrier of the scene information for few-shot action recognition, and MTFAN [69] improves the transferability of video embedding by leveraging motion patterns extracted from videos. Unlike the above methods that enhance the video representation with the help of auxiliary modality, this paper explicitly investigates the natural complementarity between different modalities based on the modality-specific posterior distributions.

Active Learning. Active learning (AL) [49] is proposed to reduce the expensive labeling cost in machine learning via acquiring most informative data from unlabeled pool for annotation. Numerous AL approaches leverage uncertainty sampling to select data points that the model produces low confidence [4,9,59]. Diversity sampling is another common method used in AL, which picks a set of typical samples via clustering [6] or core-set selection [25, 48]. Since few-shot learning shares the same goal of improving performance with limited labeled samples as AL, researchers began to consider employing AL in few-shot learning. For example, there are methods to select examples worth labeling during meta-test or meta-training by clustering approach [2, 7] or reinforcement learning [67]. Pezeshkpour et al. [45] attempt to seek the most informative samples to add into the support set during meta-test. Different from these methods, in this paper, we adopt the idea of active learning to find the more reliable modality for each sample to help the meta-training and meta-test.

Knowledge Distillation. Knowledge distillation [29, 42, 62] is proposed to distill knowledge from well-learned teacher networks to student networks, which has shown its potential in cross-modal tasks recently. For example, Gupta et al. [27] transfer supervision from labeled RGB images to unlabeled depth and optical flow images to learn rich representations with knowledge distillation. Garcia et al. [24] propose a distillation framework for action classification with a four-step process that hallucinates depth features into RGB frames. MARS [13] distills knowledge from the optical flow data to the RGB by matching high-level features and trains the network to simulate motion flow with RGB to avoid the computation of optical flow at test time. Dai et al. [14] learn an augmented RGB representation with the knowledge distilled from optical flow for action detection. Different from previous methods, we design a bidirectional knowledge distillation to actively transfer discriminative task-specific knowledge across different modalities.

3. Method

3.1. Problem Definition

In this paper, we conduct multimodal few-shot action recognition by utilizing meta-learning paradigm which consists of two stages: meta-training and meta-test. In the meta-training phase, we have a multimodal video data set $\mathbb{D}_{train}$ from base action classes C_{train}. We randomly construct multiple meta-tasks (also called episodes) from $\mathbb{D}_{train}$ to learn a meta-learner that can generalize well to novel action classes. Each meta-task $\mathcal{T}$ is comprised of a query set $\mathcal{Q} \subset \mathbb{D}_{train}$ and a support set $\mathcal{S} \subset \mathbb{D}_{train}$. In the N-way K-shot setting, the query set $\mathcal{Q} = \{(x_i^r, x_i^f, y_i)\}_{i=1}^{M}$ contains M multimodal query samples, where x_i^r and x_i^f denote two modalities (i.e., RGB and optical flow in this work) of the i^{th} query sample x_i, and $y_i \in \{1, 2, ..., N\}$ denotes the class label. The support set $\mathcal{S} = \{(x_i^r, x_i^f, y_i)\}_{i=M+1}^{M+NK}$ contains K multimodal samples for each of the N classes. In the meta-test phase, we have a multimodal test set $\mathbb{D}_{test}$ from novel classes C_{test}, and $C_{test} \cap C_{train} = \varnothing$. We construct the support and query set for each test task in a similar way as in the meta-training. Note that the class label of each query sample is invisible during meta-test. The meta-learner needs to correctly classify each sample in the query set based on only the labeled samples in the support set.

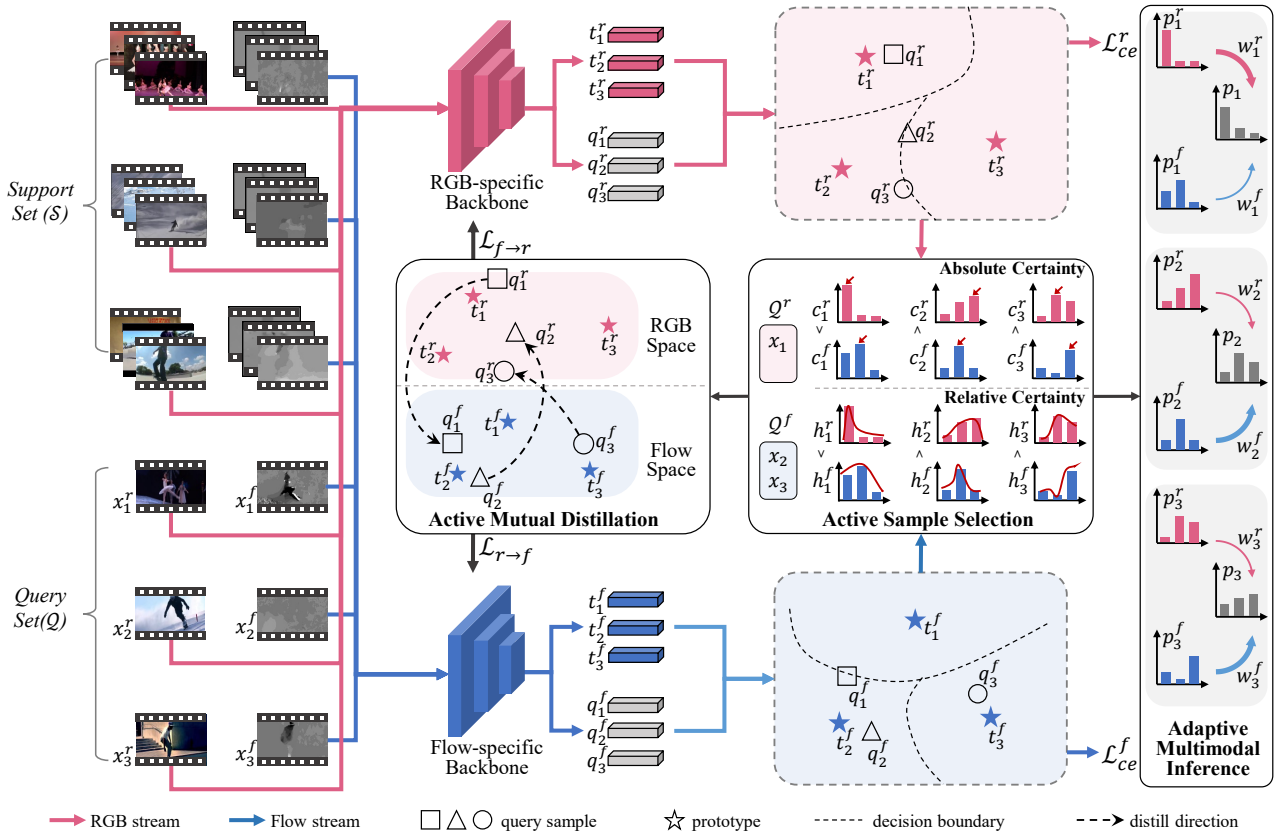

Figure 2. Illustration of the proposed AMFAR framework in the 3-way 3-shot setting. Firstly, query representations (i.e., q_i^r and q_i^f) and prototypes (i.e., t_k^r and t_k^f) are obtained from modality-specific backbone networks for both RGB and optical flow. Secondly, Active Sample Selection (ASS, Sec. 3.3) is adopted to organize query samples with large differences in the reliability of two modalities into RGB-dominant group Q^r and Flow-dominant group Q^f. Thirdly, Active Mutual Distillation (AMD, Sec. 3.4) is adopted to capture discriminative task-specific knowledge from the reliable modality to improve the representation learning of the unreliable modality. Finally, Adaptive Multimodal Inference (AMI, Sec. 3.5) is adopted to combine the predictions of different modalities by paying more attention to the reliable modality. Best viewed in color.

3.2. Overview

The overall architecture of AMFAR is illustrated in Figure 2. In each episode, we adopt a backbone network $\phi^m(\mathcal{Q}, \mathcal{S}; \theta^m)$ to obtain the representations of query samples $\{q_i^m\}_{i=1}^M$, $q_i^m \in \mathbb{R}^{d^m}$, and prototypes of support samples $\{t_k^m\}_{k=1}^N$, $t_k^m \in \mathbb{R}^{d^m}$, for each modality $m \in \{r, f\}$. Note that the modality-specific backbone networks will be elaborated in the experiment. We further compute the modality-specific posterior distribution for each query sample according to query-to-prototype distances in modality-specific feature space. In the meta-training phase, we firstly adopt **Active Sample Selection** to select query samples with large differences in the reliability of two modalities and organize them into two groups, i.e., RGB-dominant group $\mathcal{Q}^r$ and Flow-dominant group $\mathcal{Q}^f$. The RGB-dominant group contains samples where RGB is more reliable, and the Flow-dominant group is defined in the same way. The reliability of a specific modality for each sample is estimated based on the certainty of the modality-specific posterior distribution. Then, for the selected samples in $\mathcal{Q}^r$ and $\mathcal{Q}^f$, we adopt **Active Mutual Distillation** to capture discriminative task-specific knowledge from the reliable modality to enhance the representation learning of unreliable modality through a bidirectional distillation mechanism. In the meta-test phase, we adopt **Adaptive Multimodal Inference** to make the adaptive fusion decision based on modality-specific posterior distributions by paying more attention to the reliable modality.

3.3. Active Sample Selection

In this module, we select query samples with large reliability differences between the two modalities, where the more reliable modality is considered to be the sample-specific dominant modality. We define the reliable modality for each query sample as the one that can reflect more task-specific discriminative characteristic, and thus deserves more attention in few-shot learning. For a query sample, the reliable modality may be inconsistent in different tasks, because the contribution of a specific modality highly depends

on the contextual information of both query and support samples in each few-shot task. To deal with this problem, we propose to estimate the reliability of different modalities based on the certainty of modality-specific posterior distributions. In each episode, the modality-specific posterior distribution $p_i^m \in \mathbb{R}^N$ for the i^{th} query sample can be computed as:

$$\mathcal{P}(\hat{y}_i = k|x_i^m) = \frac{\exp\big(-\psi(q_i^m, t_k^m)\big)}{\sum_{k'=1}^{N}\exp\big(-\psi(q_i^m, t_{k'}^m)\big)}, \quad k \in \{1, ..., N\}, \;\; m \in \{r, f\},$$

where $q_i^m \in \mathbb{R}^{d^m}$ denotes the modality-specific representation of the i^{th} query sample and and $t_k^m \in \mathbb{R}^{d^m}$ denotes the prototype of the k^{th} class. ψ is a distance measurement function, i.e., Euclidean distance is used in this work. Inspired by uncertainty-based active learning algorithms [4, 9, 59], we consider estimating the modality reliability with two measurements: absolute certainty and relative certainty. We define the absolute certainty c_i^m as the maximum element of the modality-specific posterior distribution:

$$c_i^m = \max_k \mathcal{P}(\hat{y}_i = k|x_i^m). \tag{1}$$

In addition, we define the relative certainty h_i^m as the negative self-entropy of the modality-specific posterior distribution:

$$h_i^m = \sum_{k=1}^{N} \mathcal{P}(\hat{y}_i = k|x_i^m) \log \mathcal{P}(\hat{y}_i = k|x_i^m). \tag{2}$$

For each query sample, if a specific modality achieves high absolute certainty and relative certainty, this modality is reliable enough to express discriminative action characteristic in the few-shot task. Conversely, if the certainty is low, the modality is probably unreliable to identify actions in the few-shot task. To facilitate the exploring of cross-modal complementarity, we select query samples with large differences in the reliability of two modalities and organize them into two groups:

$$\begin{aligned} \mathcal{Q}^r &= \left\{ (x_i^r, x_i^f) \;\middle|\; (x_i^r, x_i^f) \in \mathcal{Q}, c_i^r > c_i^f, h_i^r > h_i^f \right\}, \\ \mathcal{Q}^f &= \left\{ (x_i^r, x_i^f) \;\middle|\; (x_i^r, x_i^f) \in \mathcal{Q}, c_i^f > c_i^r, h_i^f > h_i^r \right\}, \end{aligned} \tag{3}$$

where $\mathcal{Q}^r$ denotes the RGB-dominant group, that contains query samples whose RGB modality has higher certainty in the few-shot task. Analogously, $\mathcal{Q}^f$ denotes the Flow-dominant group, where Flow is more reliable to identify different query samples.

3.4. Active Mutual Distillation

In this section, we propose an active mutual distillation strategy to improve the representation learning of the unreliable modality by exploiting task-specific discriminative knowledge from the reliable modality. Before introducing the proposed active mutual distillation, we review the conventional knowledge distillation methods that utilize a well-trained teacher model to guide the learning of the student model with consistency constraint. One of the popularly used consistency constraints is KL divergence computed based on logits:

$$\mathcal{D}_{KL}(p^1, p^2) = \sum_{i=1}^{N} p_i^1 \log \frac{p_i^1}{p_i^2}, \tag{4}$$

where p^1 and p^2 denote the logits produced by the teacher and the student respectively. In existing methods, the above teacher-student distillation is conducted on individual samples consistently.

To take advantage of the complementarity between different modalities to promote the few-shot action recognition, a straightforward idea is to conduct distillation by regarding the prediction model learned on one modality as teacher and another modality as student. However, it is difficult to determine which modality should be a teacher, since the contribution of a specific modality varies on different samples and it highly depends on the context information of the few-shot task. Therefore, we propose to dynamically conduct the knowledge distillation on each sample by actively assigning the more reliable modality as the teacher. Specifically, we constrain the learning of the two modality-specific models by actively transferring query-to-prototype relation knowledge across different modalities:

$$\begin{aligned} \mathcal{L}_{f\to r}(\theta^r) &= \frac{1}{\sum\limits_{(x_i^r, x_i^f)\in\mathcal{Q}^f} c_i^f} \sum_{(x_i^r, x_i^f)\in\mathcal{Q}^f} c_i^f \mathcal{D}_{KL}(p_i^f, p_i^r), \\ \mathcal{L}_{r\to f}(\theta^f) &= \frac{1}{\sum\limits_{(x_i^r, x_i^f)\in\mathcal{Q}^r} c_i^r} \sum_{(x_i^r, x_i^f)\in\mathcal{Q}^r} c_i^r \mathcal{D}_{KL}(p_i^r, p_i^f), \end{aligned} \tag{5}$$

where p_i^r (or p_i^f) denotes the modality-specific posterior distribution for the i^{th} query sample as defined in Eq. (1). c_i^r (or c_i^f) denotes the absolute certainty defined in Eq. (1), which is used here to strengthen the distillation effect for query samples with high decision certainty.

3.5. Adaptive Multimodal Inference

In this section, we introduce how to adaptively fuse multimodal prediction results as the final decision in few-shot inference. Considering the reliability difference between the two modalities of each query sample, we design an adaptive multimodal fusion strategy:

$$\mathcal{P}(\hat{y}_i = k|x_i^r, x_i^f) = \frac{\exp\big(-w_i^r\psi(q_i^r, t_k^r) - w_i^f\psi(q_i^f, t_k^f)\big)}{\sum\limits_{k'=1}^{N}\exp\big(-w_i^r\psi(q_i^r, t_{k'}^r) - w_i^f\psi(q_i^f, t_{k'}^f)\big)}, \tag{6}$$

where ψ denotes Euclidean distance function. w_i^r and w_i^f are adaptive fusion weights for RGB and optical flow respectively. Considering that the modality-specific posterior distributions are

not always accurate in meta-test and the relative certainty cannot directly reflect the similarity between the query sample and a specific class prototype, we calculate the adaptive fusion weights with the modality-specific absolute certainties c_i^r and c_i^f:

$$w_i^r = \frac{c_i^r}{c_i^r + c_i^f}, \quad w_i^f = \frac{c_i^f}{c_i^f + c_i^r}. \tag{7}$$

3.6. Optimization

The proposed AMFAR can be optimized with the following objective function:

$$\mathcal{L} = \mathcal{L}_{ce}^r(\theta^r) + \mathcal{L}_{ce}^f(\theta^f) + \lambda\big(\mathcal{L}_{f\to r}(\theta^r) + \mathcal{L}_{r\to f}(\theta^f)\big), \tag{8}$$

where $\mathcal{L}_{f\to r}$ and $\mathcal{L}_{r\to f}$ are mutual distillation losses defined in Eq. (5). λ is balance weight. $\mathcal{L}_{ce}^r$ and $\mathcal{L}_{ce}^f$ are cross-entropy losses used to constrain the modality-specific predictions:

$$\mathcal{L}_{ce}^r(\theta^r) = CE(p_i^r, y_i), \quad \mathcal{L}_{ce}^f(\theta^f) = CE(p_i^f, y_i). \tag{9}$$

Note that the parameters of the modality-specific backbone networks, i.e., θ^r and θ^f, are updated with different losses:

$$\begin{aligned} \theta^r &\leftarrow \theta^r - \gamma\nabla_{\theta^r}\big(\mathcal{L}_{ce}^r(\theta^r) + \lambda\mathcal{L}_{f\to r}(\theta^r)\big), \\ \theta^f &\leftarrow \theta^f - \gamma\nabla_{\theta^f}\big(\mathcal{L}_{ce}^f(\theta^f) + \lambda\mathcal{L}_{r\to f}(\theta^f)\big), \end{aligned} \tag{10}$$

where γ denotes learning rate.

4. Experiments

4.1. Datasets

We evaluate our approach on four popularly used challenging few-shot action recognition benchmarks: Kinetics [10], Something-Something V2 (SSv2) [26], HMDB51 [32], and UCF101 [54]. We consistently extend each of these datasets to multimodal by generating optical flow frame sequences from the raw videos with dense optical flow algorithm [37]. For Kinetics [10] and SSv2 [26], we follow the same splits as in [74] and [8], which both randomly select 100 classes from the whole dataset with 64/12/24 classes used for training/validation/test. For HMDB51 [32] and UCF101 [54], we use the splits from [72], where the 51 classes in HMDB51 are split into 31 training classes, 10 validation classes and 10 test classes, while the 101 classes in UCF101 are split into 70/10/21 classes for training/validation/test.

4.2. Implementation Details

Data Pre-Processing. Following the pre-processing procedure used in STRM [58], we sparsely and uniformly sample 8 moments from the temporal sequence of each video. For each moment, we utilize one RGB frame as the visual data, and two consecutive optical flow frames as the motion data. The input of the multimodal few-shot action recognition task are the sequences of the sampled RGB and optical flow frames.

Modality-Specific Backbones. For RGB modality, following previous works [44, 58], we utilize ResNet-50 [28] pre-trained on ImageNet [15] as the backbone to extract frame-level visual features. For optical flow modality, we adopt I3D [10] pretrained on Charades [50] to extract the frame-level motion features. Then, following STRM [58], we obtain the video-level features for both RGB and optical flow modalities with multiple enhanced frame-level feature pairs and calculate the query-specific prototype via aggregating video-level features of support samples from the corresponding action class. The dimensions of sample features are 2048 and 1024 for the RGB and optical flow modalities respectively, i.e., $d^r = 2048$ and $d^f = 1024$.

Learning. During meta-training, following [44, 58], we resize each frame to 256×256 and randomly crop a 224×224 region for both RGB and optical flow modalities. We set the balance weight (i.e., λ defined in Eq.(8)) to 1.0 for all benchmarks. All models are trained end-to-end with SGD optimizer. Following the setting in [58], we firstly use cross-entropy loss to train RGB-specific and Flow-specific networks separately until convergence. We further optimize parameters of RGB-specific and Flow-specific networks with the objective defined in Eq.(8) for 5,000 episodes on all datasets, where the learning rate (i.e., γ) is 10^{-7}.

Evaluation. Following existing few-shot action recognition methods [8, 44, 58], we conduct 5-way 1-shot and 5-way 5-shot experiments on the four benchmarks. In meta-test, we follow prior works [8, 44, 58] to construct 10,000 episodes and report the mean accuracy. RGB and optical flow frames are directly resized to 224×224 without cropping.

4.3. Comparison with State-of-the-Art Methods

We compare the proposed AMFAR with both unimodal and multimodal methods on four benchmarks. For RGB modality, we directly compare with existing state-of-the-art few-shot action recognition methods. For optical flow, since there are no existing methods, we compare with several representative vision-based methods by retraining them on optical flow data. For multimodal baselines, besides comparing with several few-shot methods (i.e., AmeFu-Net [22] and MTFAN [69]) that utilize auxiliary modality data, we extend representative vision-based methods through early fusion, i.e., fusing multimodal features before conducting the few-shot reasoning, or late fusion (LF), i.e., fusing the prediction results of independently learned modality-specific models. We use two kinds of early fusion strategies including concatenation (EC) and Co-Attention (EA).

The comparison results for the 5-way 1-shot and 5-way 5-shot tasks are shown in Table 1. Based on the results, we have the following observations. **(1) Comparison with unimodal methods.** In the 5-way 5-shot setting, our AMFAR outperforms the state-of-the-art unimodal method HyRSM [66] on the RGB modality of SSv2 and HMDB51 by significant margins of 10.5% and 11.8% respectively. In the 5-way 5-shot setting on Kinetics, the state-of-the-art unimodal method STRM [58] has the accuracy of 86.7% and 69.7% on RGB and optical flow respectively, while our AMFAR obtains much better result of 92.6%. In the 5-way 1-shot setting, our AMFAR performs better than HyRSM [66] on the RGB modality of Kinetics and SSv2 by 6.4% and 7.4%. These significant improvements demonstrate the necessity of exploring the complementarity between different modalities in few-shot action recognition. **(2) Comparison with multimodal methods.** Although the multimodal baselines achieve remarkable performances on most benchmarks, they cannot consistently outperform all the unimodal methods. In contrast, our AMFAR performs much better than existing multimodal methods, and performs consistently better than all unimodal methods on all datasets. For ex-

Table 1. Comparison with state-of-the-art few-shot action recognition methods. We use † to mark methods that are re-implemented by ourselves. For multimodal approaches extended from existing unimodal methods, "EC" denotes the early fusion scheme of concatenation, "EA" denotes the early fusion scheme of Co-Attention, and "LF" denotes late fusion. "-" means the result is not available in published works.

Modality	Method	Kinetics		SSv2		HMDB51		UCF101	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
RGB	Matching Net [74]	53.3	78.9	-	-	-	-	-	-
	ProtoNet † [53]	55.5	84.6	26.7	53.3	45.2	71.9	70.9	94.4
	MAML [74]	54.2	78.9	-	-	-	-	-	-
	CMN [74]	60.5	78.9	-	-	-	-	-	-
	TARN [5]	66.6	78.5	-	-	-	-	-	-
	ARN [72]	63.7	82.4	-	-	44.6	59.1	62.1	84.8
	OTAM [8]	73.0	85.8	42.8	52.3	-	-	-	-
	TRX [44]	63.6	85.9	42.0	64.6	-	75.6	-	96.1
	TA2N [33]	72.8	85.8	47.6	61.0	59.7	73.9	81.9	95.1
	HyRSM [66]	73.7	86.1	54.3	69.0	60.3	76.0	83.9	94.7
	STRM [58]	-	86.7	-	68.1	-	77.3	-	96.9
Flow	ProtoNet-F [53]†	45.2	69.5	32.9	51.1	43.7	65.0	69.7	89.6
	TRX-F [44]†	44.8	69.7	30.7	52.4	43.0	67.6	65.6	90.6
	STRM-F [58]†	47.8	69.7	36.3	55.7	52.2	67.9	79.7	91.6
Multimodal	ProtoNet-EC [53]†	63.8	84.1	33.0	49.5	56.9	73.8	78.3	93.9
	ProtoNet-EA [53]†	61.7	83.9	31.1	50.5	53.2	76.3	76.7	94.3
	ProtoNet-LF [53]†	58.5	86.9	33.3	59.5	52.0	78.0	81.5	97.4
	AmeFu-Net [22]	74.1	85.8	-	-	60.2	75.5	85.1	95.5
	MTFAN [69]	74.6	87.4	45.7	60.4	59.0	74.6	84.8	95.1
	TRX-LF [44]†	65.9	86.8	37.2	61.1	57.4	78.2	81.6	94.1
	STRM-EC [58]†	68.3	87.4	45.5	66.7	59.3	78.3	87.4	96.3
	STRM-EA [58]†	68.4	87.0	44.1	62.4	60.3	76.3	85.4	94.7
	STRM-LF [58]†	66.9	87.7	41.4	70.4	55.0	81.3	83.8	98.4
	AMFAR(ours)	**80.1**	**92.6**	**61.7**	**79.5**	**73.9**	**87.8**	**91.2**	**99.0**

Table 2. Ablation results on Kinetics and SSv2.

ASS		AMD	AMI	Kinetics		SSv2	
AC	RC			1-shot	5-shot	1-shot	5-shot
✘	✔	✔	✔	72.9	89.9	57.9	73.9
✔	✘	✔	✔	77.8	89.5	58.8	78.6
✔	✔	✘	✔	77.2	90.4	55.1	78.2
✔	✔	✔	✘	72.9	89.1	50.4	73.8
✔	✔	✔	✔	80.1	92.6	61.7	79.5

ample, our AMFAR achieves large improvements of 16.0% in the 1-shot setting on the challenging SSv2 dataset compared with the second best multimodal approach, i.e., MTFAN [69]. And in the 5-shot setting on SSv2, AMFAR performs better than the second best multimodal approach STRM-LF [58] by 9.1%. In addition, AMFAR increases the performance by 11.7% and 13.6% in 1-shot setting on Kinetics and HMDB51 compared with STRM-EA [58]. These results show that directly applying existing unimodal methods cannot well solve the multimodal few-shot action recognition, and also show the importance of considering task-specific context information in exploring the cross-modal complementarity.

4.4. Ablation Study

We analyze the impact of the three key components, i.e., ASS, AMD and AMI, in our AMFAR on two challenging benchmarks including Kinetics and SSv2. Ablation results in 5-way 1-shot and 5-way 5-shot settings are shown in Table 2. Compared with our full model, removing the AMD module results in a performance drop of 2.9% and 6.6% in 1-shot setting on Kinetics and SSv2 respectively. The performance of the variant model without AMI decreases by 3.5% and 5.7% in 5-shot setting on Kinetics and SSv2 respectively. In addition, we also analyze the impact of the two modality reliability measurements in the ASS module, i.e., absolute certainty and relative certainty, which are denoted as AC and RC respectively. We observe that the performance of our approach drops from 80.1% to 72.9% or 77.8% in 5-way 1-shot setting on Kinetics, when the absolute certainty measurement or relative certainty measurement is removed. The accuracy drop is more significant when the absolute certainty measurement is removed, indicating that absolute certainty is more effective than relative certainty when only one measurement is used. The above results demonstrate the importance of three components in our method.

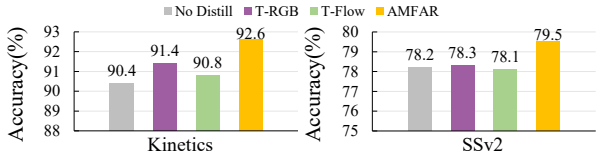

Figure 3. Comparison with conventional distillation strategies in 5-way 5-shot setting. T-RGB (or T-Flow) denotes distillation where RGB (or optical flow) is consistently regarded as teacher.

4.5. Further Remarks

Influence of Different Distillation Strategies. To further investigate the effectiveness of the proposed active mutual distillation, we study the impact of different distillation strategies on Ki-

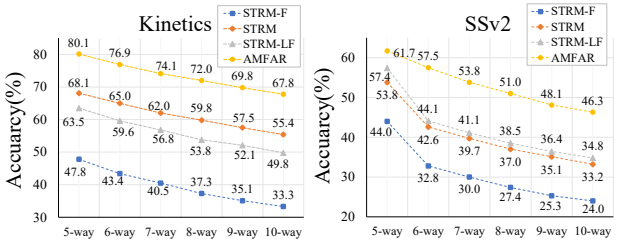

Figure 4. N-way 1-shot performance on Kinetics and SSv2.

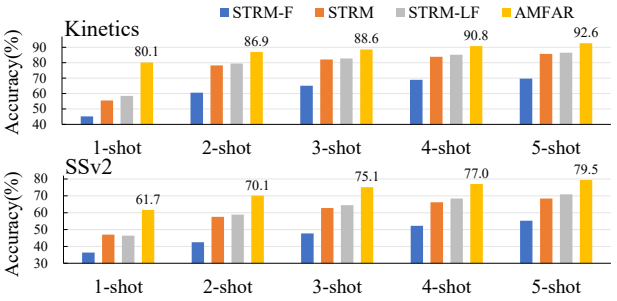

Figure 5. Comparison results with different number of support samples in 5-way K-shot setting.

netics and SSv2. We compare our AMD with two conventional distillation strategies including T-RGB and T-Flow, where the former consistently regards the RGB modality as teacher and the flow modality as student for each sample, and the later conducts the distillation in the opposite direction. As shown in Figure 3, the proposed AMD performs much better than conventional distillation strategies on two benchmarks. For example, our approach achieves the performance gain of 1.2% over the T-RGB on both Kinetics and SSv2. These results demonstrate the advantage of the proposed AMD.

N-way Few-Shot Classification. To investigate the performance of the proposed AMFAR under more challenging conditions, we show the results of our model when using more action classes on Kinetics and SSv2 in Figure 4. As shown, on the challenging SSv2 dataset, with the increase of N, our model has much larger performance gains compared with the competitive baselines, i.e., STRM [58], STRM-F [58], and STRM-LF [58], which demonstrates the generalization ability of our model.

Performance with Different Number of Support Samples. To more comprehensively analyze the performance of our model in different few-shot scenarios, we conduct extra experiment on Kinetics and SSv2 by increasing the number of support samples from 1-shot to 5-shot. As shown in Figure 5, too few support samples limit the performance of AMFAR and baselines, but AMFAR performs consistently better than baselines in all settings. For example, AMFAR outperforms the second best method by more than 5.6% on all settings of SSv2 dataset. Note that, AMFAR outperforms the second best method by a large margin of 21.6% and 15.3% in the extremely challenging few-shot setting (i.e., K=1) on Kinetics and SSv2 respectively, which further demonstrates the generalization ability of our model.

Parameter Analysis. In Figure 6, we analyze the impact of the balance weight λ in objective function (8). When λ is small (i.e., $\lambda < 0.5$), the performance of AMFAR is worse, because the contribution of the AMD module is restricted and the discriminative knowledge cannot be well distilled across different modalities. On the contrary, a too large balance weight might reduce the influence of the cross-entropy loss, leading to the slight degradation on performance. In practice, our AMFAR achieves best performance (i.e., 92.6% and 79.5%) on Kinetics and SSv2 with the balance weight of 1.0.

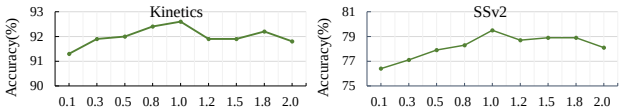

Figure 6. Parameter analysis of λ in 5-way 5-shot setting.

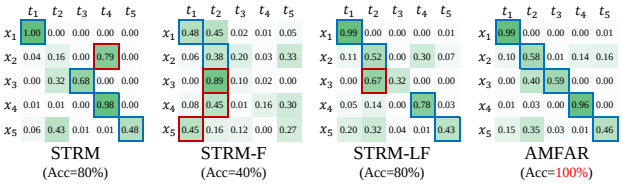

Figure 7. Visualization of the similarity between query samples (rows) and prototypes (columns) in a meta-test episode from Kinetics. The action classes, from left to right, are *shearing sheep*, *throwing axe*, *busking*, *diving cliff* and *blasting sand*. The blue box indicates correct prediction and red box indicates incorrect prediction.

4.6. Qualitative Results

To qualitatively compare our method with competitive baselines, i.e., STRM, STRM-F, and STRM-LF, we visualize the detailed query-to-prototype similarities for an episode in Figure 7. As shown, multimodal methods (i.e., STRM-LF and AMFAR) can make more accurate decisions than unimodal methods (i.e., STRM and STRM-F). Specifically, for the second query sample in Figure 7, the incorrect decision obtained by the unimodal method STRM based on the RGB modality can be rectified by fusing the results of different modalities. Additionally, the multimodal method STRM-LF produces a wrong decision for the third query sample because it cannot identify the reliable modality, while our model can produce correct result because it regards RGB as the reliable modality and avoids the negative effect of the incorrect result made based on optical flow. These results further demonstrate the effectiveness of our method.

5. Conclusion

We propose a novel Active Multimodal Few-shot Action Recognition (AMFAR) framework, which is the first attempt to apply the idea of active learning in exploring the multimodal complementarity for few-shot action recognition. The proposed AMFAR can actively find the more reliable modality based on the task-specific context information to improve the representation learning of the unreliable modality and also improve the few-shot inference in meta-test through adaptive fusion. Experiments on four public datasets demonstrate that the proposed method significantly outperforms existing unimodal and multimodal methods.

Acknowledgments. This work was supported by National Key Research and Development Program of China (2021ZD0112200), National Natural Science Foundation of China (No. 62036012, 62072455, 61721004, U20B2070), Beijing Natural Science Foundation (L201001).

References

[1] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. In *arXiv preprint arXiv:1810.09502*, 2018. 2

[2] Peyman Bateni, Jarred Barber, Raghav Goyal, Vaden Masrani, Jan-Willem van de Meent, Leonid Sigal, and Frank Wood. Beyond simple meta-learning: Multi-purpose models for multi-domain, active and continual few-shot learning. In *arXiv preprint arXiv:2201.05151*, 2022. 3

[3] Peyman Bateni, Raghav Goyal, Vaden Masrani, Frank Wood, and Leonid Sigal. Improved few-shot visual classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14493–14502, 2020. 3

[4] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9368–9377, 2018. 3, 5

[5] Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. In *arXiv preprint arXiv:1907.09021*, 2019. 1, 3, 7

[6] Zalán Bodó, Zsolt Minier, and Lehel Csató. Active learning with clustering. In *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, pages 127–139. JMLR Workshop and Conference Proceedings, 2011. 2, 3

[7] Rinu Boney and Alexander Ilin. Semi-supervised and active few-shot learning with prototypical networks. In *arXiv preprint arXiv:1711.10856*, 2017. 3

[8] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10618–10627, 2020. 1, 3, 6, 7

[9] Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Sequential graph convolutional network for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9583–9592, 2021. 3, 5

[10] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[11] Chaofan Chen, Xiaoshan Yang, Changsheng Xu, Xuhui Huang, and Zhe Ma. Eckpn: Explicit class knowledge propagation network for transductive few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6596–6605, 2021. 2

[12] Chaofan Chen, Xiaoshan Yang, Ming Yan, and Changsheng Xu. Attribute-guided dynamic routing graph network for transductive few-shot learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6259–6268, 2022. 2

[13] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2019. 3

[14] Rui Dai, Srijan Das, and François Bremond. Learning an augmented rgb representation with cross-modal knowledge distillation for action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13053–13064, 2021. 3

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[16] Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: spatially-aware few-shot transfer. In *Advances in Neural Information Processing Systems*, volume 33, pages 21981–21993, 2020. 3

[17] Xuanyi Dong, Linchao Zhu, De Zhang, Yi Yang, and Fei Wu. Fast parameter adaptation for few-shot image captioning and visual question answering. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 54–62, 2018. 3

[18] Sai Kumar Dwivedi, Vikram Gupta, Rahul Mitra, Shuaib Ahmed, and Arjun Jain. Protogan: Towards few shot learning for action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 1308–1316, 2019. 1, 3

[19] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2020. 3

[20] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1

[21] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 2

[22] Yuqian Fu, Li Zhang, Junke Wang, Yanwei Fu, and Yu-Gang Jiang. Depth guided adaptive meta-fusion network for few-shot video recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1142–1151, 2020. 1, 3, 6, 7

[23] Nuno Cruz Garcia, Sarah Adel Bargal, Vitaly Ablavsky, Pietro Morerio, Vittorio Murino, and Stan Sclaroff. Distillation multiple choice learning for multimodal action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2755–2764, 2021. 1

[24] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018. 3

[25] Yonatan Geifman and Ran El-Yaniv. Deep active learning over the long tail. In *arXiv preprint arXiv:1711.00941*, 2017. 3

[26] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim,

Valentin Haenel, Ingo Fruend, Peter Yianilos, and Moritz Mueller-Freitag. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 6

[27] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016. 3

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[29] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *arXiv preprint arXiv:1503.02531*, volume 2, 2015. 3

[30] Hanzhe Hu, Shuai Bai, Aoxue Li, Jinshi Cui, and Liwei Wang. Dense relation distillation with context-aware aggregation for few-shot object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10185–10194, 2021. 3

[31] Leonid Karlinsky, Joseph Shtok, Sivan Harary, Eli Schwartz, Amit Aides, Rogerio Feris, Raja Giryes, and Alex M Bronstein. Repmet: Representative-based metric learning for classification and few-shot object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5197–5206, 2019. 3

[32] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 6

[33] Shuyuan Li, Huabin Liu, Rui Qian, Yuxi Li, John See, Mengjuan Fei, Xiaoyuan Yu, and Weiyao Lin. Ta2n: Two-stage action alignment network for few-shot action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1404–1411, 2022. 3, 7

[34] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019. 1

[35] Weide Liu, Chi Zhang, Guosheng Lin, and Fayao Liu. Crnet: Cross-reference networks for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4165–4173, 2020. 3

[36] Yaoyao Liu, Bernt Schiele, and Qianru Sun. An ensemble of epoch-wise empirical bayes for few-shot learning. In *European Conference on Computer Vision*, pages 404–421. Springer, 2020. 2

[37] Bruce D Lucas, Takeo Kanade, et al. *An iterative image registration technique with an application to stereo vision*, volume 81. Vancouver, 1981. 6

[38] Qiguang Miao, Yunan Li, Wanli Ouyang, Zhenxin Ma, Xin Xu, Weikang Shi, and Xiaochun Cao. Multimodal gesture recognition based on the resc3d network. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3047–3055, 2017. 2

[39] Frederik Pahde, Oleksiy Ostapenko, Patrick Jä Hnichen, Tassilo Klein, and Moin Nabi. Self-paced adversarial training for multimodal few-shot learning. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 218–226. IEEE, 2019. 3

[40] Frederik Pahde, Mihai Puscas, Tassilo Klein, and Moin Nabi. Multimodal prototypical networks for few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2644–2653, 2021. 3

[41] Jinxing Pan, Xiaoshan Yang, Yi Huang, and Changsheng Xu. Few-shot egocentric multimodal activity recognition. In *ACM Multimedia Asia*, pages 1–7. 2021. 3

[42] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019. 3

[43] Fang Peng, Xiaoshan Yang, and Changsheng Xu. Sgva-clip: Semantic-guided visual adapting of vision-language models for few-shot image classification. *arXiv preprint arXiv:2211.16191*, 2022. 3

[44] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-relational crosstransformers for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 475–484, 2021. 1, 3, 6, 7

[45] Pouya Pezeshkpour, Zhengli Zhao, and Sameer Singh. On the utility of active instance selection for few-shot learning. In *NeurIPS HAMLETS*, 2020. 3

[46] A J Piergiovanni and Michael S Ryoo. Representation flow for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9945–9953, 2019. 1

[47] Viktor Rozgic, Sankaranarayanan Ananthakrishnan, Shirin Saleem, Rohit Kumar, Aravind Namandi Vembu, and Rohit Prasad. Emotion recognition using acoustic and lexical features. In *INTERSPEECH*, volume 2012, pages 366–369, 2012. 2

[48] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *arXiv preprint arXiv:1708.00489*, 2017. 3

[49] Burr Settles. Active learning literature survey. University of Wisconsin-Madison Department of Computer Sciences, 2009. 3

[50] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 6

[51] Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtash Harandi. Adaptive subspaces for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4136–4145, 2020. 2

[52] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, volume 27, 2014. 1

[53] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, volume 30, 2017. 2, 7

[54] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. In *arXiv preprint arXiv:1212.0402*, 2012. 6

[55] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2019. 2

[56] Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, and Wei Zhang. Optical flow guided feature: A fast and robust motion representation for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1390–1399, 2018. 1

[57] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H S Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018. 2

[58] Anirudh Thatipelli, Sanath Narayan, Salman Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Bernard Ghanem. Spatio-temporal relation modeling for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19958–19967, 2022. 1, 3, 6, 7, 8

[59] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. In *booktitle of machine learning research*, volume 2, pages 45–66, 2001. 3, 5

[60] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 1

[61] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S M Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 200–212, 2021. 3

[62] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1365–1374, 2019. 3

[63] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, and Daan Wierstra. Matching networks for one shot learning. In *Advances in neural information processing systems*, volume 29, 2016. 2

[64] Jun Wan, Sergio Escalera, Gholamreza Anbarjafari, Hugo Jair Escalante, Xavier Baró, Isabelle Guyon, Meysam Madadi, Juri Allik, Jelena Gorbova, and Chi Lin. Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3189–3197, 2017. 2

[65] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9197–9206, 2019. 3

[66] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Mingqian Tang, Zhengrong Zuo, Changxin Gao, Rong Jin, and Nong Sang. Hybrid relation guided set matching for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19948–19957, 2022. 1, 3, 6, 7

[67] Mark Woodward and Chelsea Finn. Active one-shot learning. In *arXiv preprint arXiv:1702.06559*, 2017. 3

[68] Jiamin Wu, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. Task-aware part mining network for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8433–8442, 2021. 2

[69] Jiamin Wu, Tianzhu Zhang, Zhe Zhang, Feng Wu, and Yongdong Zhang. Motion-modulated temporal fragment alignment network for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9151–9160, 2022. 1, 3, 6, 7

[70] Yongqin Xian, Bruno Korbar, Matthijs Douze, Lorenzo Torresani, Bernt Schiele, and Zeynep Akata. Generalized few-shot video classification with video retrieval and feature generation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE, 2021. 3

[71] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8808–8817, 2020. 2

[72] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip H S Torr, and Piotr Koniusz. Few-shot action recognition with permutation-invariant attention. In *European Conference on Computer Vision*, pages 525–542. Springer, Cham, 2020. 1, 3, 6, 7

[73] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 803–818, 2018. 1

[74] Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 751–766, 2018. 3, 6, 7