# Few-Shot Video Classification via Representation Fusion and Promotion Learning

Haifeng Xia[1]*, Kai Li[2], Martin Renqiang Min[2], Zhengming Ding[1]
[1]Department of Computer Science, Tulane University, [2]NEC Labs, America
{hxia,zding1}@tulane.edu, {kaili,renqiang}@nec-labs.com

## Abstract

*Recent few-shot video classification (FSVC) works achieve promising performance by capturing similarity across support and query samples with different temporal alignment strategies or learning discriminative features via Transformer block within each episode. However, they ignore two important issues: a) It is difficult to capture rich intrinsic action semantics from a limited number of support instances within each task. b) Redundant or irrelevant frames in videos easily weaken the positive influence of discriminative frames. To address these two issues, this paper proposes a novel Representation Fusion and Promotion Learning (RFPL) mechanism with two sub-modules: meta-action learning (MAL) and reinforced image representation (RIR). Concretely, during training stage, we perform online learning for seeking a task-shared meta-action bank to enrich task-specific action representation by injecting global knowledge. Besides, we exploit reinforcement learning to obtain the importance of each frame and refine the representation. This operation maximizes the contribution of discriminative frames to further capture the similarity of support and query samples from the same category. Our RFPL framework is highly flexible that it can be integrated with many existing FSVC methods. Extensive experiments show that RFPL significantly enhances the performance of existing FSVC models when integrated with them.*

## 1. Introduction

Due to the emergence of deep learning [43, 18, 15], video action recognition has obtained impressive improvement by learning informative semantics from raw videos [23, 38]. These achievements heavily rely on the powerful supervision of considerable well-labeled videos to optimize deep neural networks [32, 41]. However, the collection and annotation of training samples become laborious and expensive, especially for video sequences. Hence, such a learning paradigm does not always feasible for many practical applications [7, 29]. This conflict motivates investigations on few-shot video classification (FSVC) which aims to learn a model of good generalization performance from a few labeled video samples [4, 35, 37].

The mainstream solutions to FSVC typically adopt metric-based meta-learning [30, 12] fashion by training model across multiple episodes where each includes one support set with a few annotated videos and the other query set with unlabeled instances. For inference, the well-trained model measures the similarities across support and query videos within each episode to determine which categories the query samples come from [17]. However, it is difficult to precisely assess frame-wise relationships within video sequences, since their temporal information has significant divergence or even mismatch. To overcome such a challenge, the intuitive strategy is to achieve temporal alignment among cross-video frames. For example, OTAM [4] develops a variant of dynamic time wrapping (DTW) [27] to find the optimal alignment path via the frame-level cumulative distance function. Similarly, ITANet [45] explores implicit temporal alignment via a traversal strategy. Moreover, TRX [28] develops multiple tuples of sub-sequence to achieve action matching. And STRM [35] studies self-attention mechanisms to emphasize channel representation in a single video. HyRSM [37] presents a set matching concept to explore temporal alignment without frame-level ordering across various videos.

Although these existing works have achieved appealing results, they tend to encounter performance bottlenecks due to ignorance of the following two issues. First, the limited number of labeled samples within each individual task hardly provides sufficient support to assist model learning the essential action semantics for the recognition task. Second, there are redundant or irrelevant frames in videos; treating these less informative frames in the same way as the other informative ones likely weakens the discriminability of the learned video representations, and thus triggers a negative effect on solving the FSVC task.

In this paper, we propose a novel **R**epresentation **F**usion

*Work done during the internship at NEC Laboratories America.

and **P**romotion **L**earning (**RFPL**) framework to address the two issues, by two novel modules, respectively. To address the deficiency of semantic information within each individual task, we propose the meta-action learning (MAL) module which introduces external knowledge learned globally to enrich video representations learned locally within each individual task. Specifically, MAL maintains a global meta-action bank storing representations of atomic action snippets that can be used to compose various action categories of higher complexity. The global meta-action bank is shared by all action types and all individual FSVC tasks. We explore it to enrich the video feature representations learned locally in each individual task via a novel Single Value Decomposition (SVD) technique.

To address the redundant/irrelevant frame issue, we propose the reinforced image representation (RIR) module which explores reinforcement learning to discover the importance of each frame to the FSVC task. We train the reinforcement learning agent by exploiting the relationship between support and query video pairs, striving to emphasize the influence of discriminative frames to help the model accurately capture sample-wise similarity. Our MAL and RIR can be easily plugged into the existing FSVC methods to promote their performance. Our main contributions in this work are summarized in three folds as:

- We propose the meta-action learning module which learns a global meta-action bank to enrich video representation learned locally in individual tasks, via a novel SVD technique.
- We address the negative effect of task-irrelevant frames on capturing cross-video relationships and develop a reinforced image representation module that promotes the contribution of discriminative frames.
- Our RFPL framework is highly flexible such that it can be integrated with various FSVC methods and gets improved performance.

## 2. Related Works

### 2.1. Few-Shot Video Classification (FSVC)

When using meta-learning manner [8, 9], FSVC scenario assumes that each episode includes a few labeled support videos and query ones without annotations and aims to learn and generalize a model to identify unseen classes [3, 50]. The current solutions are based on metric learning. Concretely, raw video sequences are first converted into corresponding high-level feature representations with DNN. The prediction of the query sample is determined by the matching support instance with the highest similarity [48]. However, video data with temporal dimension content is extremely complicated so it is difficult to precisely measure sample-wise relationships. To eliminate such a drawback, the current works generally calculate the similarity after temporal alignment. For instance, OTAM [4] assumes that frame orderings across various videos are consistent and utilizes dynamic time wrapping to find the frame-level alignment path with a cumulative distance matrix. ITANet [45] adopts a traversal manner to match frames across different videos. TRX [28] first builds tuples of sub-sequence and exploits them to match action. STRM [35] advances TRX by considering spatial information within each video. HyRSM [38] explores set matching [13, 46] metric to efficiently compute cross-video connection. Differently, our proposed RFPL focuses on video representation enhancement by injecting meta-action knowledge and precise similarity measurement via the importance-aware refinement of discriminative frames with reinforcement learning.

### 2.2. Reinforcement Learning (RL)

Reinforcement Learning (RL) arises from the neuroscientific and psychological study that which actions humans can do to gain more benefits from the environment [16, 33]. In other words, RL is utilizing the interaction of agent and environment to learn a policy that brings the most reward [14]. In fact, such a sequential decision-making procedure can be formulated as a Markov Decision Process (MDP) [6]. Recently, due to the powerful learning ability of DNN, the advanced RL works typically deploy a trainable DNN to approximate the value function or policy function as Deep Q-Network (DQN) [26] and Deep Deterministic Policy Gradient (DDPG) [22]. In addition, many real-world applications also consider using RL to improve model performance such as object tracking [42], protein analysis [20], video summarization [21] and action recognition [34]. Unlike them, our deployment of RL is solving the few-shot video action scenario by producing frame-wise weight to promote the effect of discriminative frames. Moreover, we also conduct delicate designs on DQN to be suitable for FSVC.

## 3. Proposed Method

### 3.1. Problem Setup

In few-shot video classification (FSVC), the main task is to learn a high-generalization model to identify novel action categories of video with only a few annotated video instances. To make model training and test environment consistent, the mainstream learning strategy is meta-learning fashion [28]. Concretely, we can access a label-sufficient meta-training set $\mathcal{D}_{train}$ and another meta-testing set $\mathcal{D}_{test}$, and there is **no category overlap** across these two sets. During the training stage, videos of each episodic are collected from $\mathcal{D}_{train}$ and further divided into a support set and a query set. The support set includes $k$-samples in each of $n$ classes, which is typically defined as $n$-way-$k$-shot scenario. Under this condition, the relationship across support and query samples will be explored to identify which cate-

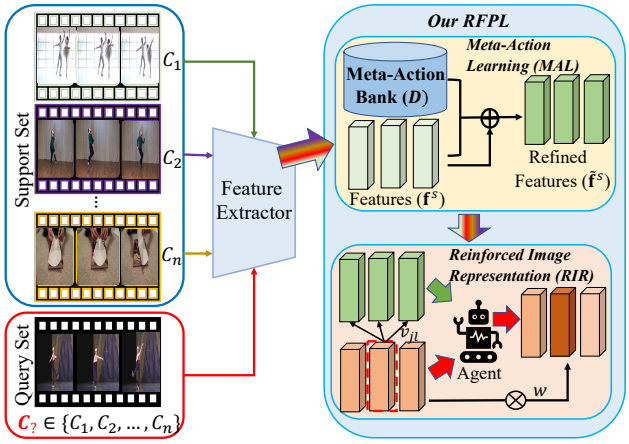



Figure 1. Overview of our representation fusion and promotion learning (RFPL) framework. RFPL involves two sub-modules: meta-action learning (MAL) fusing task-shared action semantics to enrich video representation and reinforced image representation (RIR) highlighting the effect of discriminative frames to precisely learn sample-wise similarity.

gory each query video belongs to. In terms of the test stage, each episodic also consists of support and query set which are derived from $\mathcal{D}_{test}$.

### 3.2. Framework Overview

The key to solving FSVC problem lies in how to effectively learn the pattern of discriminative action semantics from a few labeled support videos. Motivated by this, we propose a flexible **Representation Fusion & Promotion Learning** (RFPL) framework to enrich semantic information of video features via two novel sub-modules, as Fig.1 shows. One meta-action learning module utilizes episodes of meta-training set to build a task-shared action bank further refining feature embedding via knowledge fusion. Another reinforced image representation module adjusts the importance of frames to achieve the enhancement of discriminative semantics with reinforcement learning.

#### 3.2.1 Meta-Action Learning

It is acknowledged that a video can be decomposed into multiple image frames corresponding to sub-actions, i.e., $\mathbf{X}_i = \{\mathbf{x}_{ij}\}_{j=1}^m$ where $m$ is the number of frames in the $i$-th video. For example, a long jump video records some sub-actions: running, jumping and landing. In addition, we also observe that several videos come from different categories but they do share similar sub-actions such as *jumping in long jump* and *playing basketball videos*. When these two categories appear in different episodes, the transfer of similar patterns across them will facilitate the model to easily and accurately discover the similarity between support and query samples.

To achieve such a goal, the intuitive manner is to construct a task-shared action bank $\mathbf{D}$ to store representations of considerable sub-actions. Hence, the similar sub-actions in the long jump and playing basketball videos can be represented by the same atoms of this bank. Specifically, the widely-used backbone (e.g., ResNet-50) is first utilized to extract high-level feature of each frame, resulting in one video being transformed into the corresponding feature set $\mathbf{F}_i = \{\mathbf{f}_{ij} | \mathbf{f}_{ij} = \Phi(x_{ij})\}_{j=1}^m$, where $\Phi$ is the feature generator. And support videos within each episode are also converted into a feature matrix $\mathbf{F}^s \in \mathbb{R}^{d \times nkm}$ where each column denotes the feature of one frame with the dimension as $d$. To this end, we can gradually build and refine the memory bank $\mathbf{D} \in \mathbb{R}^{d \times \kappa}$ across different training episodes:

$$\min_{\mathbf{D},\mathbf{A}} \|\mathbf{F}^s - \mathbf{D}\mathbf{A}\|_{\mathrm{F}}^2 + \|\mathbf{A}\|_{\ell_1}, \tag{1}$$

where $\mathbf{D}$ is the learnable dictionary and $\mathbf{A}$ is sparse coefficient with $\ell_1$-norm, which controls the sparseness of coefficient to achieve better reconstruction by using as fewer atoms as possible and efficiently avoids information redundancy in $\mathbf{D}$. Since the current fruitful solutions to FSVC generally adopt episodic training fashion without repeatedly accessing the previous task, we expect to gradually involve action semantics per episode into task-shared bank $\mathbf{D}$, thus, the online gradient update manner [25, 24] is adopted to optimize $\mathbf{D}$ and $\mathbf{A}$ instead of closed-form solution.

After obtaining the task-shared meta-action bank, another consideration is how to adaptively fuse such semantics into each specific episode. Consequently, we rethink the meaning of learned action bank. In fact, it constitutes an action space where each atom represents a basis vector. But not all are needed for the current episode due to the assumption that the number of action categories in $\mathbf{D}$ is much larger than that in this episode. Therefore, a reasonable adaptation operation is to highlight the episode-relevant basis vectors from $\mathbf{D}$. Fortunately, the classical single value decomposition (SVD) [40, 1] helps us achieve this expectation. Specifically, we adopts SVD to decompose support feature matrix $\mathbf{F}^s$ and action memory bank $\mathbf{D}$ as:

$$\mathrm{SVD}(\mathbf{F}^s) \Rightarrow \mathbf{U}^s\mathbf{\Sigma}^s\mathbf{V}^{s\mathsf{T}}, \quad \mathrm{SVD}(\mathbf{D}) \Rightarrow \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}}, \tag{2}$$

where single values $\Sigma$ or $\Sigma^s$ control the contribution of corresponding basis vectors. So far, it is intuitive to obtain two observations from the decomposition. First, the information of feature space basis in $\mathbf{U}$ is more robust and generic than that of $\mathbf{U}^s$, since action bank refers to sufficient videos. Second, compared with $\Sigma$, $\Sigma^s$ can better estimate the intrinsic importance of basis vectors due to the fact that it is directly captured from the current episode. Based on these points, we consider fusing the advantage of each component to form the episode-adaptive action bank as $\mathbf{D}^e = \mathbf{U}\Sigma^s\mathbf{V}^{s\mathsf{T}}$. Note that $\mathbf{D}^e = \mathbf{U}(\mathbf{U}^{s\mathsf{T}}\mathbf{U}^s)\Sigma^s\mathbf{V}^{s\mathsf{T}} =$

$\mathbf{U}\mathbf{U}^{s\top}\mathbf{F}^s$, which is exactly a rotation of $\mathbf{F}^s$ using the feature basis correlation between $\mathbf{U}$ and $\mathbf{U}^s$. This rotation matrix brings more robust and generic semantic knowledge from $\mathbf{D}$ to $\mathbf{F}^s$.

The next task is fusing the robust action semantics $\mathbf{D}^e$ into video representations. Meanwhile, we discover that the inner product between frame feature $\mathbf{f}_{*j}^{s/q}$ and $\mathbf{d}_l^e \in \mathbf{D}^e$ actually reflects the projection coefficient on each space basis, which supports us to re-express video features via $\mathbf{D}^e$. The adaptive updating strategy is formalized as:

$$\widetilde{\mathbf{f}}_{*j}^{s/q} = (1-\rho)\mathbf{f}_{*j}^{s/q} + \rho \sum\nolimits_{l=1}^{\kappa} \alpha_{jl}\mathbf{d}_l^e, \tag{3}$$

where $\alpha_{jl} = \mathbf{cos}(\mathbf{f}_{*j}^{s/q}, \mathbf{d}_l^e)$ denotes the cosine similarity, $\rho$ balances accepting external knowledge and keeping original representations, and $\mathbf{d}_l^e$ is the $l$-th column basis vector of $\mathbf{D}^e$. Since support and query samples both undergo re-expression via $\mathbf{D}^e$, their similar or dissimilar associations will be further adjusted and promoted.

#### 3.2.2 Reinforced Image Representation

Our meta-action learning module enriches their latent features via the captured generic dictionary. However, not all frames in a video contain discriminative action semantics. In other words, they belong to two types: action-relevant frames and action-irrelevant ones. Promoting the contribution of the former and minimizing influence of the latter benefits the precise estimation of similarity between support and query videos. The keys to this expectation are **automatically distinguish** these two types and **adjust their ratios** in the fused representation $\widetilde{\mathbf{f}}_j^{s/q}$. Without manual annotations on the importance of frames, it is difficult for the model to learn the distinction ability. To overcome this challenge, we can break it down into a simple situation. Given the paired support and query videos from the same action, the action-relevant frames are likely to be similar or even the same, while the remaining ones tend to be different. Calculating frame-wise similarity across two videos easily discovers the discriminative frames. Their representation enhancement obviously narrows cross-video distance. For the paired samples in various classes, the weight adjustment on frames results in the reduction of cross-video similarity. In summary, these two cases indicate that frame-wise comparison and weight modification trigger the corresponding effects. This is an analogy to the basic logic of reinforcement learning adopting an action and receiving its reward.

With the analogy, we develop a reinforced image representation (RIR) module within the reinforcement learning (RL) framework to fulfill our goal. Specifically, in FSVC, the basic concepts of RL are redefined:

• **State**: the video feature at $t$ time, i.e., $s_t = \{w_j\widetilde{\mathbf{f}}_{*j}\}_{j=1}^m$ where $w_j$ initialized as 1 is the importance of the corresponding $j$-th frame;

• **Action**: $a_t$ depends on the output of agent ($p \in [0,1]$) with the *Sigmoid* function applied. When $p < 0.5$, the action $a_t = 0$ with $\mathrm{p}(a_t) = 2p$ and $w_j = w_j - \mathrm{p}(a_t)$, otherwise, $a_t = 1$ with $\mathrm{p}(a_t) = 2(p-0.5)$ and $w_j = w_j + \mathrm{p}(a_t)$. The agent in Figure 1 is implemented as one deep neural network.

• **Reward**: $r_t$ of $t$ time includes positive type and negative one. For the paired videos in the same class, $r_t$ is positive with cross-video similarity improvement, while the reduction results in the negative one. The operation is inverse for paired videos in different classes. The reward is considered as the supervision to instruct the learning of the agent. Note that we ignore superscript $s/q$ in three concepts, since they are suitable for all videos.

In reinforcement learning, the interaction procedure between agent and environment actually is a Markov decision process [20]. Concretely, given the estimated action value, the state is transferred from $s_t^{s/q}$ to $s_{t+1}^{s/q}$ according to an implicit probability distribution. During the state transition stage, the main challenge is how to decide the optimal action at the current state. In fact, the action $a_t^{s/q}$ exists in a hidden probability space which is effectively estimated by Q-learning [26, 39] strategy. Under this condition, the expectation value of actions $a_t^{s/q}$ is formally defined as $Q_t$ and the current action modifies the importance $w_j^{s/q}$ of frame $\widetilde{\mathbf{f}}_{*j}^{s/q}$ and affects the finial reward. The formulation of $Q_t$ is:

$$Q_t(s_t^{s/q}, a_t^{s/q}) = \max_{\pi} \mathbb{E}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} \cdots | \pi], \tag{4}$$

where $\gamma$ denotes the discount factor balancing the immediate reward and the future ones, and $\pi = \arg\min_{a_t^{s/q}} Q_t$ is our desired probability distribution policy of action. To estimate such a policy, the Deep-Q network (DQN) [26] as an effective tool is generally utilized to approximate $Q_t$. In a nutshell, DQN typically constructs the learnable function taking the state as input to produce the probabilities $p$ of action space. The function is always instantiated as a parameterized network.

However, such a straightforward input design of DQN fails to observe the relationship of video pairs and embed this pivotal knowledge into action selection. For FSVC, the appropriate input of DQN is supposed to involve **three different valuable contents**. **The first part** is the feature of the chosen frame $\widetilde{\mathbf{f}}_{*j}^{s/q}$ on which the estimated action is to be conducted. Emphasizing the current frame representation significantly upgrades its contribution to the prediction. **The second one** is the fused feature of the remaining frames, i.e., $\frac{1}{m-1}\sum_{l=1,l\neq j}^{m} w_l^{s/q}\widetilde{\mathbf{f}}_{*l}^{s/q}$. This information tends to affect the computation of future rewards. **The third special point** lies in measuring the cosine similarities across this frame and other frames in another video, written as $\mathbf{v}_j \in \mathbb{R}^m$ with its element $\mathbf{v}_{jl} = \mathbf{cos}(\widetilde{\mathbf{f}}_j^{s/q}, \widetilde{\mathbf{f}}_l^{q/s})$. $\mathbf{v}_j$ reflects its potentiality to promote the cross-video similarity.

In terms of the output of DQN, $a_t^{s/q}$=0 suggests that the importance of this frame should be reduced via $w_j^{s/q} = w_j^{s/q} - \mathrm{p}(a_t^{s/q})$, where $\mathrm{p}(a_t^{s/q}) = 2p$ and $p$ is the probability of action. For the other situation, weight update of frame is $w_j^{s/q} = w_j^{s/q} + \mathrm{p}(a_t^{s/q})$ with $\mathrm{p}(a_t^{s/q}) = 2(p - 0.5)$. After conducting actions, we have $w_j^{s/q} \in (0, 2)$ and it is simple to attain the corresponding reward by calculating the change of similarity of video representations $r = \mathbf{sign}\left(\mathbf{d}(\widetilde{\mathbf{f}}^s, \widetilde{\mathbf{f}}^q)|s_{t+1} - \mathbf{d}(\widetilde{\mathbf{f}}^s, \widetilde{\mathbf{f}}^q)|s_t\right)$, where $\mathbf{d}(\widetilde{\mathbf{f}}^s, \widetilde{\mathbf{f}}^q)$ means $\mathbf{cos}(\sum_{j=1}^m w_j^s \widetilde{\mathbf{f}}_{*j}^s, \sum_{j=1}^m w_j^q \widetilde{\mathbf{f}}_{*j}^q)$, and $\mathbf{sign}(\cdot)$ denotes a sign function. Finally, with the Bellman theorem [2], the Deep-Q network is optimized with:

$$\min_{\Theta} \mathbb{E}\left[r + \gamma \max_{a_{i+1}^{s/q}} Q(s_{i+1}^{s/q}, a_{i+1}^{s/q}) - Q(s_i^{s/q}, a_i^{s/q})\right]^2, \quad (5)$$

where $\Theta$ includes all the trainable network parameters of DQN. For the training stage, the snapshot of state, action, and reward at any time is stored and divided into several mini-batches with random sampling, and we adopt $\epsilon$-greedy strategy to determine action using $\pi$ with probability $\epsilon$ and use random actions with $1 - \epsilon$. For the inference stage, the well-trained DQN is frozen and infers the weight for each frame given the paired videos. To this end, video representations are further advanced with frame-wise importance.

### 3.3. Module Deployment

Our developed Representation Fusion & Promotion Learning (RFPL) framework includes two components. Meta-action learning (MAL) stores abundant action semantics via $\mathcal{D}_{train}$ and is adapted into each specific episode to enhance video representations, while Reinforced image representation (RIR) exploits deep Q-learning theory of RL to increase the representation of informative frame. Actually, it is straightforward and convenient to plug our REM into various FSVC methods to boost the model performance. In this paper, we choose three state-of-the-art FSVC models, e.g., OTAM [4], TRX [28] and STRM [35].

Concretely, original OTAM first extracts frame-level features across all support and query samples per episode, and then calculate frame-wise similarity from support and query video pair. It finally utilizes a dynamic time wrapping algorithm (DTW) to obtain the similarity of videos and uses a metric manner to optimize or deduce the category of the query. When embedding RFPL into OTAM, our MAL module is combined with the outputs of the feature extractor, and our RIR is performed before the calculation of frame-wise similarity of video pairs. Similarly, for TRX, the average-pooling features of ResNet-50 are regarded as the inputs for MAL, and the refined representations further help in the selection of discriminative frames. Different from OTAM, RIR in TRX compares the given query instances with all support ones, determines an optimal weighted query representation with the highest similarity, and feeds it into the following modules. Compared with OTAM and TRX, STRM utilizes the feature maps before the average pooling of ResNet-50 as inputs for a Transformer block and finally compresses them into frame-level vector representation as $\mathbf{F}_i^{s/q}$ where our RFPL is conducted.

## 4. Experiments

### 4.1. Experimental Settings

**Datasets.** Three popular few-shot benchmarks are evaluated including UCF101 [31], Kinetics [5] and SSv2-Full [10]. For Kinetics and SSv2-Full, we follow the data split protocol of [4, 28, 37] to build meta-training set $\mathcal{D}_{train}$, meta-validation set $\mathcal{D}_{val}$ and meta-testing set $\mathcal{D}_{test}$, which consists 64, 12 and 24 categories, respectively. For UCF101, we adopt the split and sampling manners in [44].

**Implementation Details.** Following the existing works [4, 28, 35], 8 image frames per video are randomly and uniformly selected as its raw representations and the pre-trained ResNet-50 [11] is considered as the backbone to extract high-level features from frame-level images. For the task-shared bank, we regard $\mathbf{D}$ as one learnable tensor randomly initialized and optimized by Eq. (1) with SGD optimizer (learning rate: $10^{-3} \rightarrow 10^{-5}$) in each episode during training stage and it is frozen for the test procedure. In addition, the DQN in our RIR module includes three fully-connected (FC) layers mapping the input to the number of actions, whose training is independent of other parts. Concretely, we first train DQN with several episodes by freezing the remaining networks (ResNet-50, task-shared Bank, and Transformer block), and then update the remains with the fixed DQN by other episodes. The iterative process is repeated by many times. Other training regulations are preserved as each specific method. For inference, 10,000 episodes are randomly selected from $\mathcal{D}_{test}$ to evaluate the performance of the model.

**Baselines.** Currently, HyRSM [37] and STRM [35] achieve the state-of-the-art results on most FSVC tasks, which are considered as important baselines. Besides, other excellent works such as TRN [47], CMN [48], CMN-J [49], TARN [3], OTAM [4], TTAN [19], TRX [28], ITANet [45], ARN [44], MAML [8] and MatchingNet [36] also serve as strong competitors.

### 4.2. Result Analysis

Table 1 summarizes the results of our method and other competitors on three datasets (UCF101, Kinetics, SSv2-Full) when solving various few-shot video classification tasks. According to the results, it is straightforward to achieve several important conclusions.

Table 1. Result comparison of multiple methods on few-shot video classification problem. Evaluations are conducted on the meta-testing set of UCF101, Kinetics, and SSv2-Full under 5-way scenario. The number of support videos per class varies from 1 to 5 within each episode. The highest and second results are highlighted by **bold** and underline.

| Methods | Source | Dataset | 1-shot | 2-shot | 3-shot | 4-shot | 5-shot |
|---|---|---|---|---|---|---|---|
| ARN [44] | ECCV-20 | UCF101 | 66.3 | - | - | - | 83.1 |
| TTAN [19] | ArXiv-21 | | 80.9 | - | - | - | 93.2 |
| OTAM [4] | CVPR-20 | | 79.9 | 85.5 | 87.0 | 88.3 | 88.9 |
| TRX [28] | CVPR-21 | | 78.2 | 88.9 | 92.4 | 94.1 | 96.1 |
| STRM [35] | CVPR-22 | | 78.3 | 88.9 | 91.9 | 93.5 | **96.9** |
| HyRSM [37] | CVPR-22 | | 83.9 | – | 93.0 | – | 94.7 |
| Ours+OTAM | - | | **84.3**$\uparrow_{4.4}$ | 88.9$\uparrow_{3.4}$ | 90.2$\uparrow_{3.2}$ | 91.9$\uparrow_{3.6}$ | 92.1$\uparrow_{3.2}$ |
| Ours+TRX | - | | 82.5$\uparrow_{4.2}$ | **91.1**$\uparrow_{2.2}$ | **94.1**$\uparrow_{1.7}$ | 95.6$\uparrow_{1.5}$ | 96.3$\uparrow_{0.2}$ |
| Ours+STRM | - | | 79.7$\uparrow_{1.4}$ | 90.1$\uparrow_{1.2}$ | 93.8$\uparrow_{1.9}$ | **95.8**$\uparrow_{2.3}$ | 96.7$\downarrow_{0.2}$ |
| MatchNet [36] | NeuIPS-16 | Kinetics | 53.3 | 64.3 | 69.2 | 71.8 | 74.6 |
| MAML [8] | ICML-17 | | 54.2 | 65.5 | 70.0 | 72.1 | 75.3 |
| Plain CMN [48] | ECCV-18 | | 57.3 | 67.5 | 72.5 | 74.7 | 76.0 |
| CMN-J [49] | TPAMI-20 | | 60.5 | 70.0 | 75.6 | 77.3 | 78.9 |
| TARN [3] | BMVC-19 | | 64.8 | - | - | - | 78.5 |
| ARN [44] | ECCV-20 | | 63.7 | - | - | - | 82.4 |
| ITANet [45] | IJCAI-21 | | 73.6 | - | - | - | 84.3 |
| OTAM [4] | CVPR-20 | | 73.0 | 75.9 | 78.7 | 81.9 | 85.8 |
| TRX [28] | CVPR-21 | | 63.6 | 76.2 | 81.8 | 83.4 | 85.9 |
| STRM [35] | CVPR-22 | | 62.1 | 78.6 | 82.2 | 82.8 | 86.7 |
| HyRSM [37] | CVPR-22 | | 73.7 | 80.0 | 83.5 | 84.6 | 86.1 |
| Ours+OTAM | - | | **74.6**$\uparrow_{1.6}$ | 80.0$\uparrow_{4.1}$ | 82.1$\uparrow_{3.4}$ | 84.1$\uparrow_{2.2}$ | 86.8$\uparrow_{1.0}$ |
| Ours+TRX | - | | 66.2$\uparrow_{2.6}$ | 77.5$\uparrow_{1.3}$ | 83.8$\uparrow_{2.0}$ | **85.1**$\uparrow_{1.7}$ | 87.3$\uparrow_{1.4}$ |
| Ours+STRM | - | | 64.9$\uparrow_{2.8}$ | **80.3**$\uparrow_{1.7}$ | **84.3**$\uparrow_{2.1}$ | 84.8$\uparrow_{2.0}$ | **87.7**$\uparrow_{1.0}$ |
| CMN++ [48] | ECCV-18 | SSv2-Full | 34.4 | - | - | - | 43.8 |
| TRN++ [47] | ECCV-18 | | 38.6 | - | - | - | 48.9 |
| TTAN [19] | ArXiv-21 | | 46.3 | 52.5 | 57.3 | 59.3 | 60.4 |
| ITANet [45] | IJCAI-21 | | 49.2 | 55.5 | 59.1 | 61.0 | 62.3 |
| OTAM [4] | CVPR-20 | | 42.6 | 49.1 | 51.5 | 52.0 | 52.3 |
| TRX [28] | CVPR-21 | | 42.0 | 53.1 | 57.6 | 61.1 | 64.6 |
| STRM [35] | CVPR-22 | | 42.1 | 53.8 | 59.3 | 64.2 | 68.1 |
| HyRSM [37] | CVPR-22 | | **54.3** | **62.2** | **65.1** | **67.9** | 69.0 |
| Ours+OTAM | - | | 47.0$\uparrow_{4.2}$ | 54.6$\uparrow_{5.5}$ | 58.3$\uparrow_{6.8}$ | 60.3$\uparrow_{8.3}$ | 61.0$\uparrow_{8.7}$ |
| Ours+TRX | - | | 44.6$\uparrow_{2.6}$ | 54.2$\uparrow_{1.1}$ | 59.9$\uparrow_{2.3}$ | 63.1$\uparrow_{2.1}$ | 64.6$\uparrow_{2.3}$ |
| Ours+STRM | - | | 45.7$\uparrow_{3.6}$ | 55.8$\uparrow_{2.0}$ | 61.7$\uparrow_{2.4}$ | 66.8$\uparrow_{2.6}$ | **69.5**$\uparrow_{1.4}$ |

**First**, when plugging our proposed modules into the existing FSVC works such as OTAM, its performance gains significant improvement. Concretely, for 5-way-5-shot task on SSv2-Full, the classification accuracy of OTAM using our module attains 61.0% which is much higher than its original 52.3%. In addition, the integration of our designed modules and OTAM exceeds the original one by 4.4% on 1-shot task of UCF101. These phenomenons illustrate that our method effectively facilitates OTAM to conduct more accurate temporal alignment across support and query videos by refining video representations with task-shared knowledge and discriminative frame selection. **Second**, compared with OTAM, original TRX and STRM not only consider learning temporal relationships but also draw assistance from Transformer block to achieve information complementary within each episode. Although their methods seem to be perfect and promising, our RFPL still facilitates them to boost 2% on most tasks. The main reason for success lies in that our method further helps them enrich the representations of action semantics by introducing meta-action semantics and accurately capture the intrinsic similarity across support and query instances via the increasing effect of important image frames. **Finally**, injecting our RFPL into other baselines achieves comparable performance with HyRSM on most tasks. Especially, STRM with our RFPL surpasses HyRSM by an obvious gain margin, i.e., 1.6% on 5-shot task of Kinetics, validating that our proposed method is an efficient auxiliary tool for addressing the FSVC problem.

### 4.3. Performance Study

**Path Visualization & Confusion Matrix.** From Table 1, we conclude that our RFPL effectively promotes the perfor-

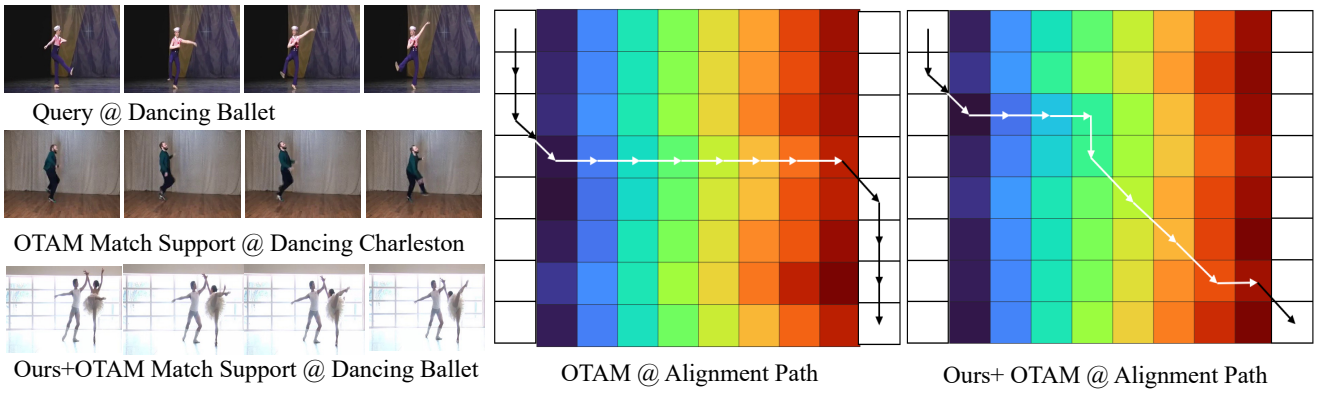
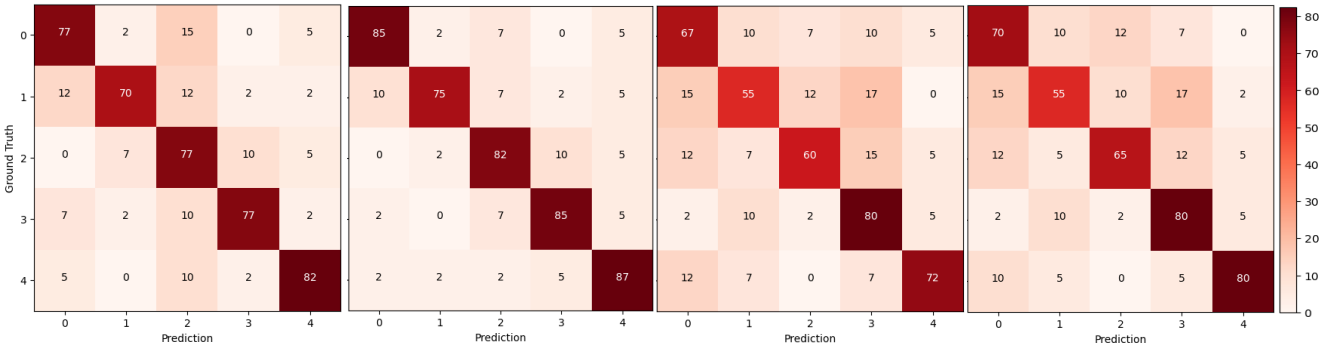



Figure 2. Comparison between plain OTAM and Ours+OTAM on action recognition result and temporal alignment path. This episode is randomly selected from the meta-testing set of Kinetics. The query video belongs to dancing ballet. Our+OTAM successfully matches it with the support sample from the same category, while plain OTAM matches into a dancing charleston video.




(a) OTAM @ Kinetics (b) Ours+OTAM @ Kinetics (c) OTAM @ SSV-Full (d) Ours+OTAM @ SSV-Full

Figure 3. Comparison of confusion matrix deduced from the model prediction of plain OTAM and Ours+OTAM. These two episodes are collected from the meta-testing set of Kinetics (2-shot) and SSv2-Full (5-shot).

mance of OTAM by a large margin. To explicitly explain the reason for improvement, we randomly select an episode from meta-testing set $\mathcal{D}_{test}$ and visualize the frame-level cumulative distance across query video and the selected support ones as Fig. 2 shows. This visualization indicates that OTAM is sensitive to image background and its temporal alignment becomes invalid. Thus, it considers that this query video from the dancing ballet class has the highest similarity with the support sample from dancing charleston. Differently, our method assists OTAM to focus on the similarity of action semantics, especially on distinguishing body posture. Hence, Ours+OTAM accurately matches the query sample with the support video from the identical category and achieves better temporal alignment.

In addition, we draw the confusion matrix of two specific episodes to further discuss their differences. Concretely, during the inference stage, for 2-shot of Kinetics or 5-shot of SSv2-Full, we randomly choose an episode with 200 query samples uniformly distributed in five categories and compare model prediction and ground-truth to form Fig. 3. For the task of Kinetics, with the help of our RFPL, the recognition abilities of OTAM in five categories are all strengthened. With respect to the more challenging task of SSv2-Full, OTAM with our method learns more discriminative action semantics and achieves significant performance improvement on several categories without reduction on the remaining ones.

**Reinforced Image Frames.** In our RFPL, reinforced image representation as one important module aims to discover the discriminative image frames via the comparison of the paired videos. To analyze its effect, we record the frame-wise weights deduced from the agent as Fig. 4 in one test episode. From two "folding paper" videos, several image frames without any "folding" action are provided with lower weights such as frames in the red box. For action-relevant images, the learned network produces higher weights for them as the blue box shows. From the other example "cutting watermelon", we have an interesting finding. The well-trained network does not give larger weights for frames including "cutting" action, while highlighting images with two halves of the red pulp of a watermelon as the green box shows. It suggests that the network considers that these image contents are useful for the improvement of cross-video similarity.

**Parameter Analysis.** There is one hyper-parameter $\rho$ in our RFPL, which serves as a trade-off between accepting external global knowledge and preserving current feature representations. In practical experiments, this parameter is

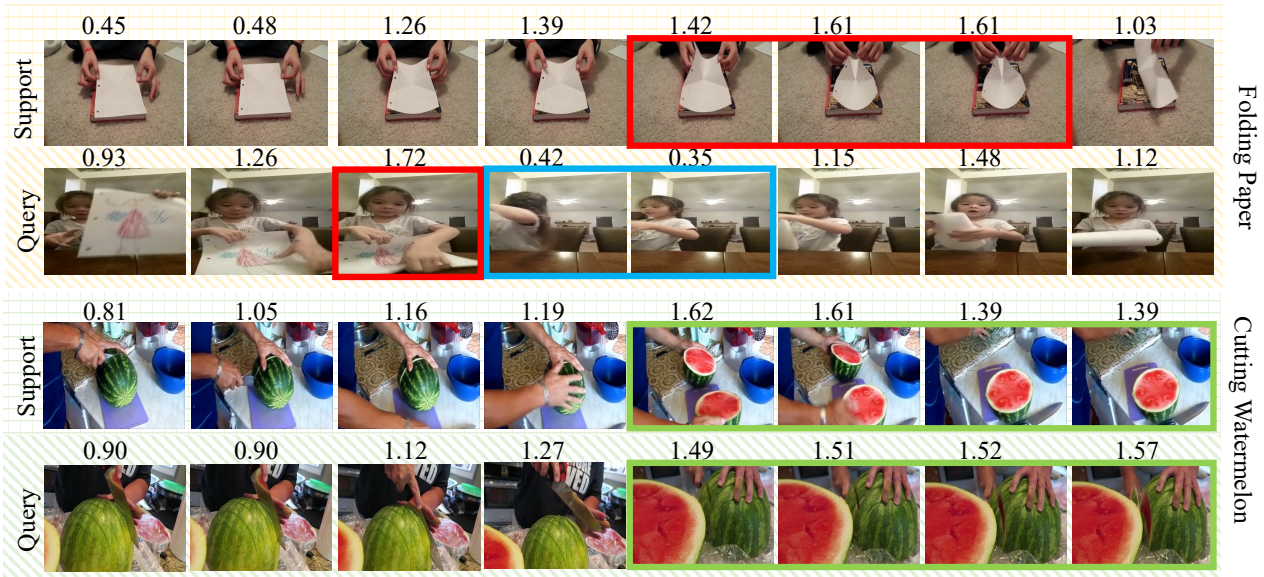



Figure 4. Importance of frames produced by reinforced image representation module. Videos of the first and second rows belong to folding paper, while the remaining videos are from cutting watermelon. Images in the red box are regarded as key action semantics by the network, while the effect of others in the blue box is reduced. Similarly, the representations of images in the green box are also enhanced.

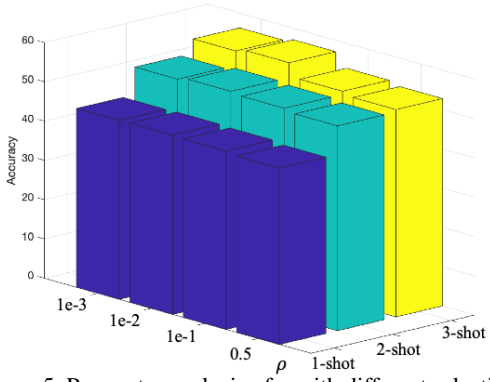



Figure 5. Parameter analysis of $\rho$ with different selections.

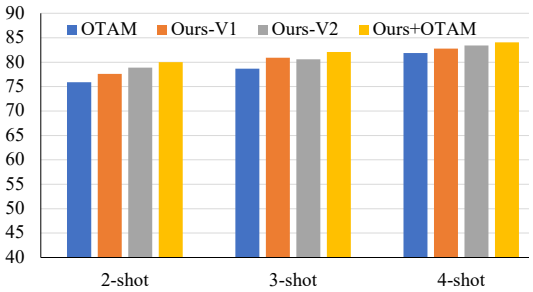



Figure 6. Ablation study. Ours-v1 and Ours-v2 mean that meta-action learning and reinforced image representation modules are plugged into the existing FSVC works, respectively.

tuned for different tasks, i.e., $\rho \in \{0.001, 0.01, 0.1, 0.5\}$. Definitely, the optimal parameter is determined by the performance of the learned model on the meta-validation set. Concretely, Fig. 5 shows the change of model classification ability on the validation set of SSv2-Full with the varying $\rho$. Based on this parameter analysis, we find that when $\rho = 0.5$, the video classification accuracy is lower than that of smaller $\rho$ on many tasks. Therefore, it is reasonable to infer that while introducing global knowledge to enhance semantics information, the protection of current information is very necessary and stabilizes the model training.

**Ablation Study.** Our RFPL consists of two modules: meta-action learning and reinforced image representation. The first one aims to enrich video representation by fusing global knowledge of the meta-action bank to each specific episode, while the second one expects to emphasize the representation of informative image frames via reinforcement learning selection. In fact, we also separately plug each module into the existing baseline, named as Ours-v1 (adding MAL) and Ours-v2 (adding RIR). These variants are combined with OTAM and evaluated on three tasks of Kinetics with results in Fig. 6. The comparisons with plain OTAM mean that each module can independently make positive contributions on advancing it. And the collaboration of MAL and RIR will bring more benefits to OTAM than each independent module. Thus, using the complete RFPL is the optimal strategy.

## 5. Conclusion

In this paper, we propose a flexible representation fusion and enhancement learning (RFPL) mechanism with two sub-modules to solve few-shot video classification. The meta-action learning module builds a generic action bank and fuses the adapted one to enrich video features, while the reinforced image representation module identifies important frames via an agent and promotes their positive effect to precisely capture cross-video similarity. Considerable experimental results and analysis verify that our RFPL effectively advances the existing FSVC methods.

## References


[1] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006. 3

[2] Richard Bellman. Dynamic programming and lagrange multipliers. *Proceedings of the National Academy of Sciences*, 42(10):767–769, 1956. 5

[3] Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. *arXiv preprint arXiv:1907.09021*, 2019. 2, 5, 6

[4] 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, 2, 5, 6

[5] 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. 5

[6] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021. 2

[7] 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

[8] 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, 5, 6

[9] Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *International Conference on Machine Learning*, pages 1920–1930. PMLR, 2019. 2

[10] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. 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. 5

[11] 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. 5

[12] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021. 1

[13] Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 1993. 2

[14] Julian Ibarz, Jie Tan, Chelsea Finn, Mrinal Kalakrishnan, Peter Pastor, and Sergey Levine. How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research*, 40(4-5):698–721, 2021. 2

[15] Taotao Jing, Haifeng Xia, Jihun Hamm, and Zhengming Ding. Marginalized augmented few-shot domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 1

[16] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996. 2

[17] Orit Kliper-Gross, Tal Hassner, and Lior Wolf. One shot similarity metric learning for action recognition. In *International Workshop on Similarity-Based Pattern Recognition*, pages 31–45. Springer, 2011. 1

[18] Kai Li, Yulun Zhang, Kunpeng Li, and Yun Fu. Adversarial feature hallucination networks for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13470–13479, 2020. 1

[19] Shuyuan Li, Huabin Liu, Rui Qian, Yuxi Li, John See, Mengjuan Fei, Xiaoyuan Yu, and Weiyao Lin. Ttan: Two-stage temporal alignment network for few-shot action recognition. *arXiv e-prints*, pages arXiv–2107, 2021. 5, 6

[20] Tingfeng Li, Shaobo Han, Martin Renqiang Min, and Dimitris N Metaxas. Learning transferable reward for query object localization with policy adaptation. *arXiv preprint arXiv:2202.12403*, 2022. 2, 4

[21] Yandong Li, Liqiang Wang, Tianbao Yang, and Boqing Gong. How local is the local diversity? reinforcing sequential determinantal point processes with dynamic ground sets for supervised video summarization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 151–167, 2018. 2

[22] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015. 2

[23] Xin Liu, Silvia L Pintea, Fatemeh Karimi Nejadasl, Olaf Booij, and Jan C van Gemert. No frame left behind: Full video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14892–14901, 2021. 1

[24] Cewu Lu, Jiaping Shi, and Jiaya Jia. Online robust dictionary learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 415–422, 2013. 3

[25] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pages 689–696, 2009. 3

[26] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015. 2, 4

[27] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007. 1

[28] 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, 2, 5, 6

[29] Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, and Abir Das. Semi-supervised action recognition with temporal contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10389–10399, 2021. 1

[30] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 1

[31] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[32] Jonathan Stroud, David Ross, Chen Sun, Jia Deng, and Rahul Sukthankar. D3d: Distilled 3d networks for video action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 625–634, 2020. 1

[33] Richard S Sutton, Andrew G Barto, et al. Reinforcement learning. *Journal of Cognitive Neuroscience*, 11(1):126–134, 1999. 2

[34] Yansong Tang, Yi Tian, Jiwen Lu, Peiyang Li, and Jie Zhou. Deep progressive reinforcement learning for skeleton-based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5323–5332, 2018. 2

[35] 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, 2, 5, 6

[36] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016. 5, 6

[37] 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, 5, 6

[38] Zhengwei Wang, Qi She, and Aljosa Smolic. Action-net: Multipath excitation for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13214–13223, 2021. 1, 2

[39] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992. 4

[40] Aming Wu, Suqi Zhao, Cheng Deng, and Wei Liu. Generalized and discriminative few-shot object detection via svd-dictionary enhancement. *Advances in Neural Information Processing Systems*, 34:6353–6364, 2021. 3

[41] Haifeng Xia, Pu Wang, and Zhengming Ding. Incomplete multi-view domain adaptation via channel enhancement and knowledge transfer. In *European Conference on Computer Vision*, pages 200–217. Springer, 2022. 1

[42] Sangdoo Yun, Jongwon Choi, Youngjoon Yoo, Kimin Yun, and Jin Young Choi. Action-decision networks for visual tracking with deep reinforcement learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2711–2720, 2017. 2

[43] Hao Zhang, Yanbin Hao, and Chong-Wah Ngo. Token shift transformer for video classification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 917–925, 2021. 1

[44] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip HS Torr, and Piotr Koniusz. Few-shot action recognition with permutation-invariant attention. In *European Conference on Computer Vision*, pages 525–542. Springer, 2020. 5, 6

[45] Songyang Zhang, Jiale Zhou, and Xuming He. Learning implicit temporal alignment for few-shot video classification. *arXiv preprint arXiv:2105.04823*, 2021. 1, 2, 5, 6

[46] Chunjiang Zhao, Wenkang Shi, and Yong Deng. A new hausdorff distance for image matching. *Pattern Recognition Letters*, 26(5):581–586, 2005. 2

[47] 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. 5, 6

[48] 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. 2, 5, 6

[49] Linchao Zhu and Yi Yang. Label independent memory for semi-supervised few-shot video classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):273–285, 2020. 5, 6

[50] Zhenxi Zhu, Limin Wang, Sheng Guo, and Gangshan Wu. A closer look at few-shot video classification: A new baseline and benchmark. *arXiv preprint arXiv:2110.12358*, 2021. 2