Task-Agnostic Self-Distillation for Few-Shot Action Recognition

Bin Zhang[1] , **Yuanjie Dang**[1,*] , **Peng Chen**[1] , **Ronghua Liang**[1] , **Nan Gao**[1] , **Ruohong Huan**[1] and **Xiaofei He**[2]

[1]Zhejiang University of Technology
[2]Zhejiang University
{201906062226, dangyj, chenpeng, rhliang, gaonan, huanrh}@zjut.edu.cn, xiaofeihe@cad.zju.edu.cn

Abstract

Task-oriented matching is one of the core aspects of few-shot Action Recognition. Most previous works leverage the metric features within the support and query sets of individual tasks, without considering the metric information across different matching tasks. This oversight represents a significant limitation in this task. Specifically, the task-specific metric feature can decrease the generalization ability and ignore the general matching feature applicable across different tasks. To address these challenges, we propose a novel meta-distillation framework for few-shot action recognition that learns the task-agnostic metric features and generalizes them to different tasks. First, to extract the task-agnostic metric information, we design a task-based self-distillation framework to learn the metric features from the training process progressively. Additionally, to enable the model with fine-grained matching capabilities, we design a multi-dimensional distillation module that extracts more detailed relations from the temporal, spatial, and channel dimensions within video pairs and improves the representative performance of metric features for each individual task. After that, the few-shot predictions can be obtained by feeding the embedded task-agnostic metric features to a common feature matcher. Extensive experimental results on standard datasets demonstrate our method's superior performance compared to existing state-of-the-art methods.

1 Introduction

Few-shot action recognition aims to learn new unseen actions with only a few data samples and thus has gained increasing attention[Perrett *et al.*, 2021; Li *et al.*, 2022]. It is promising to decrease the effort required for collecting extensive training data and address the issue of decreased accuracy that numerous effective action recognition models[Feichtenhofer *et al.*, 2019; Feichtenhofer, 2020; Arnab *et al.*, 2021; Selva *et al.*, 2023] encounter when presented with unfamiliar real-life situations. However, few-shot action recognition remains a challenging task that requires models to generalize to many unseen video tasks with only a limited number of labeled samples. Existing approaches adopt metric-based learning paradigms[Vinyals *et al.*, 2016; Snell *et al.*, 2017; Sung *et al.*, 2018]. These paradigms enable the learning of video correlation features from multiple individual tasks, which consist of support and query videos. However, these learning frameworks have limitations. They utilize task-oriented models that hinder the sharing of relative metric features among diverse tasks. Additionally, the reliance on task-specific metric features can decrease the models' generalization ability and overlook the general matching feature embedding applicable across different tasks. Some methods[Wang *et al.*, 2023a] employ large-scale data pre-trained models to enhance feature performance, which contradicts the few-shot learning problem. To address these problems, entropy-based unbiased initial model[Jamal and Qi, 2019] and deep transformer-based distillation[Xu *et al.*, 2022] have been proposed.

Entropy-based unbiased initial model[Jamal and Qi, 2019] learns an unbiased initial model with the largest uncertainty over the output labels by preventing it from over-performing in classification tasks, this meta-based method is challenging to capture complex visual information[Li *et al.*, 2023]. Deep transformer-based distillation utilizes relation distillation to align the attention distributions between the pre-trained teacher and sampled student subnetworks[Xu *et al.*, 2022], which requires massive parameters and computational resources. Recently, Self-distillation has achieved stunning results in Domain Adaptation[Yoon *et al.*, 2022; Cardace *et al.*, 2023] and self-supervised representation learning[Caron *et al.*, 2021; Song *et al.*, 2023]. These methods utilize momentum updating networks to acquire more comprehensive and universally applicable representations of consistency knowledge from a substantial volume of training data[Caron *et al.*, 2021; Zhou *et al.*, 2021]. However, in few-shot action recognition, the intricate multi-dimensional semantics of action instances may result in a decline in the student model's ability to learn metric features from the teacher model for each individual task. There is a crucial need for a global perspective to learn universal metric features across multiple episodes. Therefore, designing a reasonable distillation strategy to take full advantage of self-distillation to acquire few-shot metric features from historical episodes throughout the training process remains a challenge.

To address this issue, we introduce momentum-based knowledge distillation[Caron *et al.*, 2021] into the task-oriented training paradigm to enable the learning of general task-agnostic metric features from historical tasks, thereby enhancing the generalization ability of few-shot action classification on new tasks. Specifically, we propose a novel task-agnostic self-distillation framework for few-shot action recognition, along with a multi-dimensional distillation module. In general, our study revolves around two key ideas. First, we learn the task-agnostic metric feature by the momentum update self-distillation framework. Second, we enforce the representative performance of metric features for each individual task by a multi-dimensional distillation module. Notably, to the best of our knowledge, our method is the first to task agnostic metric feature compared to existing few-shot action recognition work that typically uses the task-oriented training paradigm.

In summary, our contribution can be summarized as follows:

- we propose a novel meta-distillation framework for few-shot action recognition that enables the model to learn task-agnostic metric features and generalize to different tasks by ensemble and summarizing the training process.
- we propose an efficient relation feature that uses a multi-dimensional distillation module to represent video metric information that enables the network to align video action instances from temporal, spatial, and channel perspectives.
- We conduct extensive experiments on three benchmark datasets to verify the effectiveness of the proposed method. The experimental results demonstrate the superior performance of our method compared to existing state-of-the-art methods.

2 Related Works

2.1 Few-Shot Action Recognition

Few-shot action recognition refers to the process of training a model capable of recognizing new classes with a limited number of labeled samples. Unlike few-shot image classification, few-shot action recognition focuses on the alignment of the temporal features in videos. The mainstream few-shot action recognition methods employ a metric-based meta-learning paradigm[Vinyals *et al.*, 2016] and perform frame-to-frame temporal matching[Cao *et al.*, 2020; Wang *et al.*, 2022; Wang *et al.*, 2023b] to search for the most similar categories. These methods require temporal or spatial matching to predict the query labels. For instance, OTAM[Cao *et al.*, 2020] enforces the alignment of video frames between query and support videos in the temporal dimension. ITANet[Zhang *et al.*, 2021] introduces a decomposed self-attention mechanism to alleviate intra-class variability in video features. TRX[Perrett *et al.*, 2021] utilizes Cross Transformers[Doersch *et al.*, 2020] to construct query-specific prototype representations. Following the matching method of TRX, STRM[Thatipelli *et al.*, 2022] enhances local and global features to effectively capture spatio-temporal contextual information in videos. MoLo[Wang *et al.*, 2023b] introduces a long-short contrastive loss to enforce local frame feature prediction with global context and perceives motion details through frame-wise difference reconstruction. SloshNet[Xing *et al.*, 2023] adaptively integrates spatial features from different levels and integrates long-term and short-term temporal features for rich spatio-temporal characteristics. Besides, some research has utilized the task-level information in few-shot action recognition, for example, HyRSM[Wang *et al.*, 2022] introduces task-aware video relation learning to customize task-specific features and utilizes a set-matching metric. However, the above researches focus on frame-level spatio-temporal relationships between videos from a single task, without fully leveraging multi-dimensional information across different tasks which may cause a decrease in the model's generalization capability on low-shot scenarios.

2.2 Self Distillation

The self-distillation scheme is a specific method within knowledge distillation that is widely applied in domain adaptation and self-supervised learning. [Hinton *et al.*, 2015] first propose Knowledge Distillation (KD) technique for training lightweight models to achieve comparable performance to deep models, which is achieved by compressing the informative knowledge from a large model (i.e., teacher model) to a small model (i.e., student model). Unlike previous knowledge distillation approaches[Hu *et al.*, 2022], the roles of the student and teacher networks in self-distillation are dynamic throughout the iterative training process. This suggests that the student and teacher networks can interchange roles, or the student networks can learn from themselves. Teacher Free Knowledge Distillation (Tf-KD)[Yuan *et al.*, 2020] showcases Knowledge Distillation (KD) as a form of label smoothing regularization, implying that label smoothing regularization can be seen as a virtual teacher, building upon these insights, Teacher-free Knowledge Distillation is proposed, allowing a student to acquire knowledge from itself or a manually-designed regularization distribution. Online subclass knowledge distillation (OSKD)[Tzelepi *et al.*, 2021] uncovers similarities within each class to grasp shared semantic information among subclasses. Throughout the online distillation process, every sample gradually converges toward representations of the same subclass while simultaneously moving away from representations of different subclasses. Semi-supervised domain adaptation (SSDA)[Yoon *et al.*, 2022] adapts a model to the target domain using self-distillation with sample pairs and generates an assistant feature by transferring an intermediate style between the teacher and the student. DINO[Caron *et al.*, 2021] introduces self-distillation into self-supervised learning by creating various distorted views or crops of an image using a multi-crop strategy, the student network processes all these crops, whereas only the global views are processed by the teacher network, thereby reinforcing ”local-to-global” correspondences. iBOT[Zhou *et al.*, 2021] performs momentum-based self-distillation on masked patch tokens and takes the teacher network as the online tokenizer, along with self-distillation on the class token to acquire visual semantics. Our method is inspired

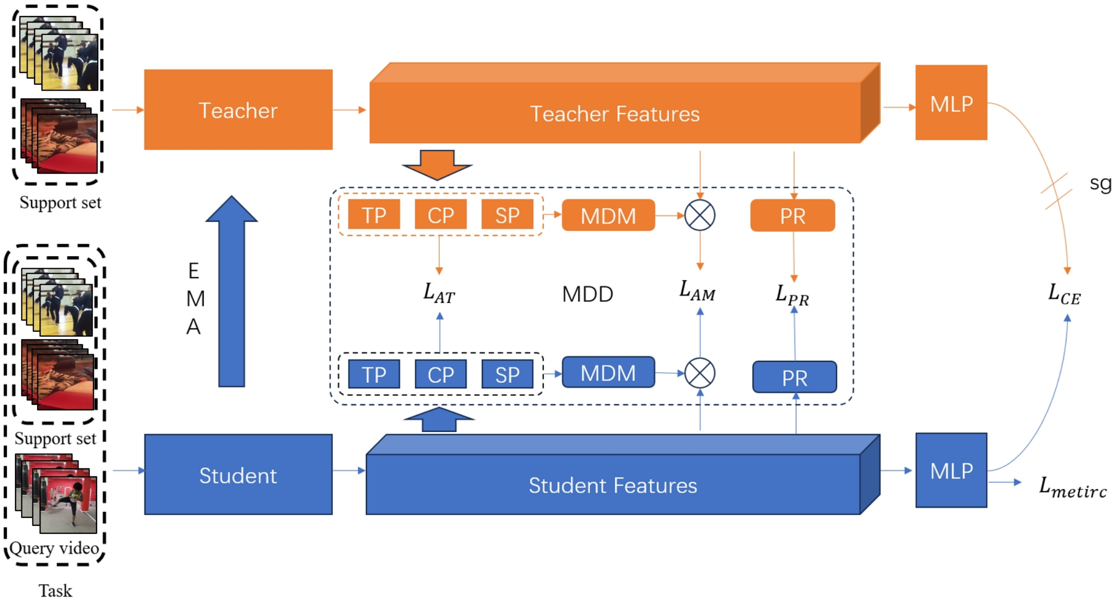


Figure 1: Illustration of our Task Agnostic Self-Distillation for Few-Shot Action Recognition. The input of the teacher is the support video set, the input of the student is the support set and query set, and the parameters of the teacher network which is visualized in orange are Exponentially Moving Averaged (EMA) updated by the student network which is visualized in blue, Multi Dimension Distillation (MDD) includes temporal Pooling (TP), Channel Pooling (CP), and Spatial Pooling (SP) to create a Multi Dimension Mask (MDM) for distillation, Pixel Relation (PR) module capture global pixel relations, ”sg” denotes stop gradient.

by momentum-based self-distillation [Caron *et al.*, 2021; Zhou *et al.*, 2021] method. Through momentum updates, the teacher network iterative ensembles the training process, enabling the extraction of task-agnostic metric features from the training process.

3 Method

3.1 Self Distillation for Few-Shot Action Recognition

Following the knowledge distillation method[Gou *et al.*, 2021; Hu *et al.*, 2022], our framework comprises two networks: the student network $\mathcal{S}$ and the teacher network $\mathcal{T}$. Unlike traditional approaches that use pre-trained networks with complex parameters as teachers, we adopted a self-distillation approach[Yuan *et al.*, 2020]. This method implies that both the student and teacher networks share identical network structures and initial parameters. To enhance the networks' ability for task-agnostic semantic metric features in episodic, few-shot learning across various tasks, we implemented a momentum-based update method inspired by DINO[Caron *et al.*, 2021]. This updated approach retains the updated network parameters within each task, the update process can be formulated as:

$$\theta_t = (1-\lambda)\cdot\theta_t + \lambda\cdot\theta_s \tag{1}$$

Where λ denotes the coefficient for momentum update, while θ_t and θ_s denote the network parameters of the teacher and student models, respectively. Unlike in self-supervised image classification, where different data augmentations are applied to the same image to create local and global views, our method in few-shot action recognition divides data into support and query sets. To ensure the student network's metric branch accurately learns metric features among different videos within the same task, we input different images of the same class from the support and query sets into the student network for learning. Simultaneously, the support set is fed into the teacher network. Leveraging the features extracted from the teacher network for the support set features (A_s) and the query set features (A_q) from the student network through a projection layer, we map them into hyperspace and train the network using cross-entropy loss as KD loss. Especially, taking the N-way 1-shot task as an example, within each task, videos are divided into support and query sets. The support set $S = \{s_1, s_2, ..., s_N\}$ comprises N action categories, each containing one video, where $s_i \in R^{T,C,H,W}$, and T denotes the number of sparsely sampled video frames used to obtain the video representation.

Given a query set video $q \in R^{T,C,H,W}$, the process begins by extracting frame-level features $F_S^{\mathcal{S}}$ for the support set S and frame-level features $F_q^{\mathcal{S}}$ for the query set video q using the student backbone network $\mathcal{S}$. Subsequently, the teacher backbone network $\mathcal{T}$ extracts frame-level features $F_S^{\mathcal{T}}$ for the support set S. We calculate the similarity between video features among $F_S^{\mathcal{T}}$ shares the same class as $F_q^{\mathcal{S}}$ by cross-entropy loss after a multi-layer projection network, this process can be formulated as:

$$L_{CE} = -\sum_{i}^{C} F_i^{\mathcal{T}s} \log F_i^{\mathcal{S}q} \tag{2}$$

Where $F_i^{\mathcal{T}s}$ denotes projected feature of $F_s^{\mathcal{T}}$ on the ith channel, $F_i^{\mathcal{S}q}$ denotes projected feature of $F_q^{\mathcal{S}}$ on the ith channel.

Further, $A_S^{\mathcal{S}}$ and $A_q^{\mathcal{S}}$ are passed into the metric module to acquire metric features among videos. These metric features guide the gradient updates within the network for the current task. Following the gradient update, the parameters of the current student network are passed to the teacher network through momentum updates.

3.2 Multi Dimension Distillation

Building upon the Attention-guided Distillation method[Zhang and Ma, 2023], we propose a multi-dimensional Attention-guided Distillation that focuses on the intricate multi-dimensional information inherent in video data. This aims to enhance the learning process of the self-distillation method mentioned in section 3.1 for few-shot action recognition, allowing for a better understanding of information across spatial, channel, and temporal dimensions within videos. We use $A \in R^{C,T,H,W}$ to denote the feature of the backbone in an action recognition model, where C, T, H, W denotes its channel number, frame number, height, and width, respectively. Then, the generation of the spatial attention map, channel attention map, and temporal attention map is equivalent to finding the mapping function, $Gs : R^{C,T,H,W} \rightarrow R^{H,W}$, $Gc : R^{C,T,H,W} \rightarrow R^{C}$ and $Gt : R^{C,T,H,W} \rightarrow R^{T}$, respectively.

Given that the absolute value of each element in the feature conveys its significance, we create $\mathcal{G}^s$ by computing the average of absolute values across the channel and temporal dimensions. Additionally, $\mathcal{G}^c$ is constructed by averaging the absolute values across the width, height, and temporal dimension, while $\mathcal{G}^t$ is formed by averaging the absolute values across the width, height, and channel dimensions. This process can be mathematically expressed as follows:

$$\mathcal{G}^c(A) = \frac{1}{THW} \sum_{T}^{i=1} \sum_{H}^{j=1} \sum_{W}^{k=1} |A_{\cdot,i,j,k}| \tag{3}$$

$$\mathcal{G}^s(A) = \frac{1}{CT} \sum_{C}^{l=1} \sum_{T}^{i=1} |A_{l,i,\cdot,\cdot}| \tag{4}$$

$$\mathcal{G}^t(A) = \frac{1}{CHW} \sum_{C}^{l=1} \sum_{H}^{j=1} \sum_{W}^{k=1} |A_{l,\cdot,j,k}| \,. \tag{5}$$

Suppose we denote l, i, j, k as the indices referring to the lth, ith, jth, and kth slice of A in the channel, temporal, height, and width dimensions respectively, the spatial attention mask M_s, the temporal attention mask M_t and the channel attention mask M_c used in attention-guided distillation are obtained by summing the attention maps derived from both the teacher and student detectors. This process can be formulated as:

$$M^s = HW \cdot \operatorname{softmax}\left(\left(\mathcal{G}^s\left(A^{\mathcal{S}}\right) + \mathcal{G}^s\left(A^{\mathcal{T}}\right)\right)/T\right) \tag{6}$$

$$M^t = T \cdot \operatorname{softmax}\left(\left(\mathcal{G}^t\left(A^{\mathcal{S}}\right) + \mathcal{G}^c\left(A^{\mathcal{T}}\right)\right)/T\right) \tag{7}$$

$$M^c = C \cdot \operatorname{softmax}\left(\left(\mathcal{G}^c\left(A^{\mathcal{S}}\right) + \mathcal{G}^c\left(A^{\mathcal{T}}\right)\right)/T\right) . \tag{8}$$

Where the superscripts $\mathcal{S}$ and $\mathcal{T}$ are used to differentiate between the student ($\mathcal{S}$) and teacher ($\mathcal{T}$) attention maps. The hyper-parameter T in the softmax function is introduced by [Hinton *et al.*, 2015] to regulate the distribution of elements in the attention masks. The attention-guided distillation loss (L_{AGD}) comprises two sub-modules: attention transfer loss (L_{AT}) and attention masked loss (L_{AM}). L_{AT} aims to prompt the student model to imitate the multi-dimension attention patterns of the teacher model. This can be formulated as:

$$\begin{aligned} L_{AT} =& L_2\left(\mathcal{G}^s\left(A^{\mathcal{S}}\right), \mathcal{G}^s\left(A^{\mathcal{T}}\right)\right) + L_2\left(\mathcal{G}^c\left(A^{\mathcal{S}}\right), \mathcal{G}^c\left(A^{\mathcal{T}}\right)\right) \\ &+ L_2\left(\mathcal{G}^t\left(A^{\mathcal{S}}\right), \mathcal{G}^t\left(A^{\mathcal{T}}\right)\right) \end{aligned} \tag{9}$$

On the other hand, L_{AM} aims to encourage the student to imitate the features of the teacher model using an L2 norm loss that is masked by M_s, M_c and M_t. This can be formulated as:

$$L_{AM} = \left((A^{\mathcal{T}} - A^{\mathcal{S}})^2 \cdot (M^s \cdot M^c \cdot M^t)\right)^{\frac{1}{2}} \tag{10}$$

3.3 Global Pixel Relation Distillation

Action recognition requires high attention to the foreground subject's motion, The relation between foreground and background features is addressed to tackle the feature imbalance issue. The Graph Convolution module GloRe[Duan *et al.*, 2019], is employed to efficiently capture global pixel relations. GloRe outperforms attention mechanisms by better capturing global context[Ni *et al.*, 2023], leading to superior distillation effects. The procedure involves extracting frame-level features separately from the teacher and student backbone networks. These features are then fed into distinct GloRe modules to grasp global pixel relations. Following this, pixel-wise relation features undergo distillation to transfer global relations from the teacher to the student, the distillation loss function can be formulated as:

$$L_{PR} = \frac{1}{k} \sum_{i=1}^{k} \left\| \phi\left(t_i\right) - f\left(\phi\left(s_i\right)\right) \right\|_2 \tag{11}$$

where k is the number of frames. t_i and s_i refer to the feature of the teacher and student respectively. ϕ represents the GloRe module.f represents adaptive convolution. The GloRe module contains three parts: graph embedding, graph convolution, and reprojection.We first embed the input signal frame feature $A \in R^{C,H,W}$ into a low-dimensional graph feature space $\bar{A} \in R^{C_1,HW}$ the graph node features $N \in R^{C_1,C_2}$ are obtained by projecting using a learnable projection matrix $M \in R^{HW,C_2}$. Then a graph convolution is applied on graph node features N to capture the relationships features $Z \in R^{C_1,C_2}$ between nodes. Finally, the global pixel relation features $F \in R^{C_1,C_2}$ are projected back into the coordinate feature space by a learnable projection matrix $M \in R^{C_2,HW}$. In addition, an adaptive convolutional mechanism is integrated into the student model to minimize feature disparities between the student and teacher models. These node features aggregate information from various regions and emphasize significant relations via a relation filter module.

4 Experiments

4.1 Datasets and Implementation Details

Datasets

We evaluate our approach on three standard datasets, including Kinetics[Carreira and Zisserman, 2017],

Method	Kinetics		UCF101		HMDB51	
	1 -shot	5-shot	1 -shot	5-shot	1 -shot	5-shot
ARN[Zhang *et al.*, 2020]	63.7	82.4	62.1	84.8	44.6	59.1
OTAM[Cao *et al.*, 2020]	72.2	84.2	79.9	88.9	54.5	68.0
ITANet[Zhang *et al.*, 2021]	73.6	84.3	-	-	-	-
TRX[Perrett *et al.*, 2021]	63.6	85.9	78.2	96.1	53.1	75.6
TA^{2}N[Li *et al.*, 2022]	72.8	85.8	81.9	95.1	59.7	73.9
STRM[Thatipelli *et al.*, 2022]	62.9	86.7	80.5	96.9	52.3	77.3
HyRSM[Wang *et al.*, 2022]	73.7	86.1	83.9	94.7	60.3	76.0
BiMHW[Thatipelli *et al.*, 2022]	72.3	84.5	81.7	89.3	58.3	69.0
HCL[Zheng *et al.*, 2022]	73.7	85.8	82.5	93.9	59.1	76.3
SloshNet[Xing *et al.*, 2023]	70.4	87.0	86.0	**97.1**	59.4	**77.5**
MoLo[Wang *et al.*, 2023b]	74.0	85.6	86.0	95.5	60.8	77.4
Ours	**76.5**	**87.4**	**88.2**	96.1	**61.1**	77.4

Table 1: Comparison with state-of-the-art few-shot action recognition methods on the Kinetics, UCF101 and HMDB51 datasets. Experiments are performed under the 5-way task with 1-shot and 5-shot settings. The best results are denoted in bold black, the second-best results are indicated with an underscore, and ”-” signifies that the result is not available in the published works.

UCF101[Soomro *et al.*, 2012], and HMDB51[Kuehne *et al.*, 2011]. The datasets are partitioned into meta-training, meta-validation, and meta-testing sets based on action categories to meet the requirements of the few-shot classification setting. For Kinetics, we follow the splitting strategy proposed by[Zhu and Yang, 2018], selecting 100 action categories, each with 100 samples, and dividing these categories into 64, 12, and 24 for training, validation, and testing, respectively. For UCF101, we split it into 70, 10, and 21 categories for training, validation, and testing. In the case of HMDB51, we split it into 31, 10, and 10 categories for training, validation, and testing, adhering to the same splitting strategy as in[Zhang *et al.*, 2020].

Implementation Details

Following the common paradigm of existing few-shot action recognition methods[Cao *et al.*, 2020; Wang *et al.*, 2023b], we employ ResNet50[He *et al.*, 2016] as the backbone network and initialize it with weights pre-trained on ImageNet[Deng *et al.*, 2009] to extract frame-level features. We sparsely and uniformly sample 8 frames from each video, like previous methods[Cao *et al.*, 2020; Wang *et al.*, 2023b]. In the network architecture, the Transformer layers of the Encoder are configured with four layers, The teacher and student models adopt the same structure and initialization parameters. During training, we resize each frame in the video into 256 × 256, followed by random horizontal flips and random cropping to a 224 × 224 region. In the testing phase, we first perform resizing and then replace random cropping with center cropping to standardize the shape of input videos of varying sizes. We utilize the Adam optimizer with an initial learning rate of 0.0005 to train our model. We randomly sample 30,000 episodes from the meta-training set for training. For testing, similar to prior work[Wang *et al.*, 2023b], we collect 10,000 episodes from the meta-testing set to evaluate the model's performance and report the average accuracy. We implement our framework using PyTorch and conduct training on one RTX 4090 GPU.

Method	SD	PR	MDD	accuracy
Baseline				86.31
Baseline + SD	✓			87.02
Baseline + SD + PR	✓	✓		87.72
Ours	✓	✓	✓	**88.32**

Table 2: Ablation study of three network components on UCF101 dataset under 5-way 1-shot settings. SD: Self-Distillation framework; PR: Pixel Relation module; MDD: Multi Dimension Distillation module

4.2 Comparison with State-of-the-Art

In the 5-way task with 1-shot, and 5-shot settings, we compare our method with state-of-the-art approaches, as presented in Table 1. Under the 1-shot settings, our method surpasses existing approaches across all three datasets. Specifically, in the 1-shot setting, our method achieves significant improvements of 2.4%, 1.8%, and 0.3% on Kinetics, UCF1O1, and HMDB51, respectively. Our method's performance in the 5-shot setting on the Kinetics dataset exceeds the previous best method by 0.4%. However, on UCF101 and HMDB51, it lags behind the TRX-based methods(e.g. STRM[Thatipelli *et al.*, 2022] and SloshNet[Xing *et al.*, 2023]).Nevertheless, our method outperforms other methods in low-shot scenarios which could attributed to the reason that low-shot scenarios contain fewer task-specific features, and our method provides extra task-agnostic information learned from the training process to assist metrics in low-shot scenarios. Therefore, the task-agnostic metric features extracted by our approach can significantly improve the recognition performance of action instances with relatively low recognition accuracy in low-shot scenarios . This problem is elaborated in the subsequent experiments.

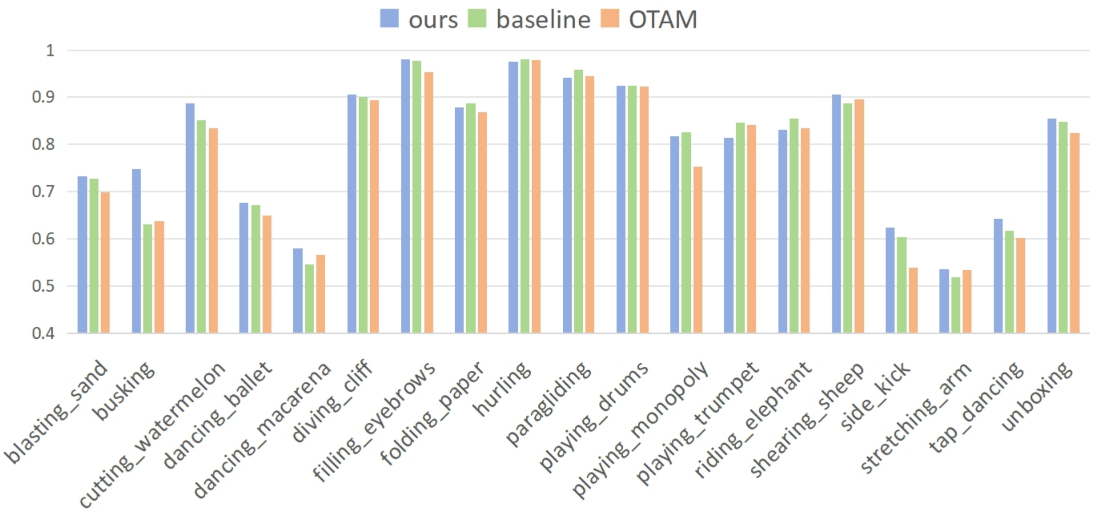


Figure 2: Comparison of the per-class meta-testing accuracy of our method with OTAM and baseline under the 5-way 1-shot setting using the Kinetics dataset.

4.3 Ablation Study

Impact of Network Components

In this subsection, we analyze the effects of three modules on the few-shot action recognition performance, Table 2 compares the performance of our method with three variants of our methods. Specifically, according to the framework characteristics and previous few-shot action recognition works[Wang *et al.*, 2022; Wang *et al.*, 2023b], we train a 3D backbone network[Tran *et al.*, 2018] with a metric module[Thatipelli *et al.*, 2022] as the ”Baseline” for ablation study, ”Baseline + SD” employs Self-Distillation(SD) framework on the baseline, consisting of two networks with the same architecture and a cross-entropy (CE) loss, ”Baseline + SD + PR” employs Pixel Relation(PR) distillation module on ”Baseline + SD”, and our method(denoted as ”Ours”) further employs Multi Dimension Distillation on ”Baseline + SD + PR”. Starting with a comparison between the ”Baseline” and ”Baseline + SD”, we observe performance improvements of 0.71% for 1-shot tasks on the UCF101 dataset, indicating that the introduction of the self-distillation framework can effectively enhance the performance of few-shot recognition.

Compared with ”Baseline + SD”, ”Baseline + SD + PR” integrates the Pixel Relation(PR) module, and an additional performance improvement of 0.7% is observed for 1-shot tasks on the UCF101 dataset, demonstrating that the pixel relation module that enhances the perception of the video’s action subject can effectively improve the extraction of few-shot action.

Subsequently, our method(Ours) further incorporates Multi-Dimension Distillation (MDD) to the ”Baseline + SD + PR”, MDD brings a significant improvement of 0.6%. This suggests that multi-dimensional information enables the self-distillation framework to perceive action metric semantics more specifically within independent tasks, ultimately enhancing overall task-agnostic metric performance. Further details on the learned metric features specifics in the teacher and student will be provided in Section 4.4.

video pair	best episode	accuracy
SS	10,000	75.68
QS	10,000	75.67
CS	**25,000**	**76.04**

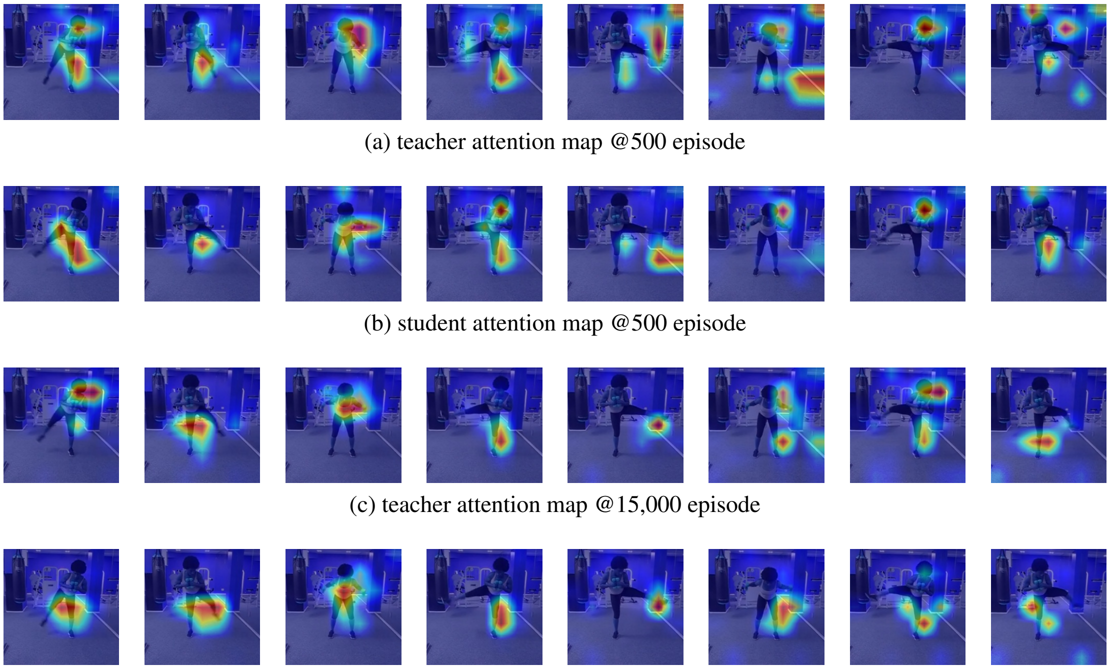
Table 3: Comparison experiments on the performance and overfit situation of different input of teacher and student network on the Kinetics datasets. ”SS” denotes the matching between support sets, ”QS” denotes the matching between query sets, and ”CS” denotes the matching between the cross set combined with support outputs from the teacher and the query set outputs from the student

Analysis of Per-Class Accuracy

To investigate the impact of our proposed method on per-class classification accuracy for specific categories, we conduct a per-class accuracy analysis on the meta-testing set of the Kinetics dataset under the 5-way 1-shot setting. A comparison is made with OTAM[Cao *et al.*, 2020] and the Baseline, following the previous few-shot action recognition works[Wang *et al.*, 2023b; Xing *et al.*, 2023], the baseline comprises a ResNet50 backbone followed by 4 Transformer layers, and BiMHW[Thatipelli *et al.*, 2022] is employed for metric learning. As illustrated in Figure 2, our method consistently exhibits superior overall performance compared to the baseline approach, emphasizing its broad applicability across various motion patterns. Notably, our approach manifests a substantial enhancement in accuracy for numerous categories characterized by initially low recognition accuracy. The results of the two previous experiments imply that the task-agnostic metric feature can significantly enhance the recognition performance of relatively low recognition accuracy action instances at low-shot scenarios.

Analysis of Different Distillation Video Pairs

To better evaluate the impact of different distillation targets within the self-distillation framework on task-agnostic metric performance, we conducted a comparison within the distilla-



(a) teacher attention map @500 episode

(b) student attention map @500 episode

(c) teacher attention map @15,000 episode

(d) student attention map @15,000 episode

Figure 3: visualization results of the student and teacher networks at episodes 500 and 15,000 on "side kick" action instance

tion branch between the teacher and student models for different input video pairs from the query and support sets. Table 3 compares the performance on three variants of input video pair. Specifically, "SS" denotes that the input videos of the student network and teacher network are both the same videos from the support set, "QS" denotes that the input videos of the student network and teacher network are both the same videos from the query set and "CS" denotes that the input videos of students are from query set but the input videos of teacher are from the support set that share the same categories with the student input. The term "best episode" refers to the point during training when the meta-test performance is optimal, and overfitting begins thereafter.

In Table 3, both "SS" and "QS" reach their optimal accuracy at the 10,000th episode, achieving 75.68% and 75.67%, respectively. However, overfitting occurred after the 10,000th episode. On the other hand, "CS" attains its best performance of 76.04% at the 25,000th episode, showing a 0.36% improvement compared to "SS" and "QS".The comparison results indicate that distilling the different videos with the same categories enables the model to learn general task-agnostic metric features, ultimately improving the generalization capability.

4.4 Visualization Results

To further investigate the disparities in the extracted metric features between student and teacher networks, we visualize attention maps for the "side kick" action instance in the Kinetics dataset. We analyze and compare the feature content of the student and teacher networks at episodes 500 and 15,000. As depicted in Figure 3, at the initial 500 episodes, the attention regions of the teacher network (a) and the student network (b) are mainly similar, with attention areas primarily focused on the subject's body and inclined towards the motion regions of the legs and arms. At the 15,000th episode, significant disparities emerge in the attention regions of the teacher (c) and student networks (d). The teacher network becomes more concentrated on features that are more universal and task-agnostic within the overall action, guiding the student network to learn more during the training phase through the self-distillation loss. In contrast, the student network focuses on more specific action metric features from multidimensional information in the current task . These outcomes concurrently affirm the findings from Section 4.3: Our task-agnostic self-distillation framework facilitates the teacher network in learning more general task-agnostic features from the learning process and guides the student network to focus on the primary subject of motion.

5 Conclusion

In this paper, we propose a novel task-agnostic self-distillation framework for few-shot action recognition. Our approach learns the general task-agnostic metric feature by distilling the learning process, thereby enhancing the generalization ability of few-shot action classification on new tasks. Additionally, we employ a Multi Dimension Distillation to enforce the representational performance of metric features for each individual task. Extensive experiments demonstrate that our method effectively extracts task-agnostic metric features by the proposed framework. Consequently, our method exhibits excellent performance and holds a performance advantage over existing few-shot action recognition methods.

Acknowledgments

*Corresponding author: Yuanjie Dang

This work was supported in part by the Natural Science Foundation of China under Grant U1909203, 62036009, 62206250, 62276237, the Zhejiang Provincial Natural Science Foundation of China under Grant LQ22F020007, LDT23F0202, LDT23F02021F02, LD24F020005, and the Ten Thousand Talent Program of Zhejiang Province.

References

[Arnab *et al.*, 2021] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[Cao *et al.*, 2020] 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.

[Cardace *et al.*, 2023] Adriano Cardace, Riccardo Spezialetti, Pierluigi Zama Ramirez, Samuele Salti, and Luigi Di Stefano. Self-distillation for unsupervised 3d domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4166–4177, 2023.

[Caron *et al.*, 2021] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[Carreira and Zisserman, 2017] 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.

[Deng *et al.*, 2009] 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.

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

[Duan *et al.*, 2019] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019.

[Feichtenhofer *et al.*, 2019] 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.

[Feichtenhofer, 2020] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020.

[Gou *et al.*, 2021] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

[He *et al.*, 2016] 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.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Hu *et al.*, 2022] Chengming Hu, Xuan Li, Dan Liu, Xi Chen, Ju Wang, and Xue Liu. Teacher-student architecture for knowledge learning: A survey. *arXiv preprint arXiv:2210.17332*, 2022.

[Jamal and Qi, 2019] Muhammad Abdullah Jamal and Guo-Jun Qi. Task agnostic meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11719–11727, 2019.

[Kuehne *et al.*, 2011] 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.

[Li *et al.*, 2022] 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.

[Li *et al.*, 2023] Wenbin Li, Ziyi Wang, Xuesong Yang, Chuanqi Dong, Pinzhuo Tian, Tiexin Qin, Jing Huo, Yinghuan Shi, Lei Wang, Yang Gao, et al. Libfewshot: A comprehensive library for few-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Ni *et al.*, 2023] Zhenliang Ni, Fukui Yang, Shengzhao Wen, and Gang Zhang. Dual relation knowledge distillation for object detection. *arXiv preprint arXiv:2302.05637*, 2023.

[Perrett *et al.*, 2021] 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.

[Selva *et al.*, 2023] Javier Selva, Anders S Johansen, Sergio Escalera, Kamal Nasrollahi, Thomas B Moeslund, and Albert Clapés. Video transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

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

[Song *et al.*, 2023] Kaiyou Song, Jin Xie, Shan Zhang, and Zimeng Luo. Multi-mode online knowledge distillation for self-supervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11848–11857, 2023.

[Soomro *et al.*, 2012] 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.

[Sung *et al.*, 2018] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS 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.

[Thatipelli *et al.*, 2022] 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.

[Tran *et al.*, 2018] 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.

[Tzelepi *et al.*, 2021] Maria Tzelepi, Nikolaos Passalis, and Anastasios Tefas. Online subclass knowledge distillation. *Expert Systems with Applications*, 181:115132, 2021.

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

[Wang *et al.*, 2022] 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.

[Wang *et al.*, 2023a] Xiang Wang, Shiwei Zhang, Jun Cen, Changxin Gao, Yingya Zhang, Deli Zhao, and Nong Sang. Clip-guided prototype modulating for few-shot action recognition. *International Journal of Computer Vision*, pages 1–14, 2023.

[Wang *et al.*, 2023b] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Changxin Gao, Yingya Zhang, Deli Zhao, and Nong Sang. Molo: Motion-augmented long-short contrastive learning for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18011–18021, 2023.

[Xing *et al.*, 2023] Jiazheng Xing, Mengmeng Wang, Yong Liu, and Boyu Mu. Revisiting the spatial and temporal modeling for few-shot action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3001–3009, 2023.

[Xu *et al.*, 2022] Dongkuan DK Xu, Subhabrata Mukherjee, Xiaodong Liu, Debadeepta Dey, Wenhui Wang, Xiang Zhang, Ahmed Awadallah, and Jianfeng Gao. Few-shot task-agnostic neural architecture search for distilling large language models. *Advances in Neural Information Processing Systems*, 35:28644–28656, 2022.

[Yoon *et al.*, 2022] Jeongbeen Yoon, Dahyun Kang, and Minsu Cho. Semi-supervised domain adaptation via sample-to-sample self-distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1978–1987, 2022.

[Yuan *et al.*, 2020] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3903–3911, 2020.

[Zhang and Ma, 2023] Linfeng Zhang and Kaisheng Ma. Structured knowledge distillation for accurate and efficient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Zhang *et al.*, 2020] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip HS Torr, and Piotr Koniusz. Few-shot action recognition with permutation-invariant attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 525–542. Springer, 2020.

[Zhang *et al.*, 2021] Songyang Zhang, Jiale Zhou, and Xuming He. Learning implicit temporal alignment for few-shot video classification. In *IJCAI*, 2021.

[Zheng *et al.*, 2022] Sipeng Zheng, Shizhe Chen, and Qin Jin. Few-shot action recognition with hierarchical matching and contrastive learning. In *European Conference on Computer Vision*, pages 297–313. Springer, 2022.

[Zhou *et al.*, 2021] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

[Zhu and Yang, 2018] 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.