Knowledge-Based Systems 304 (2024) 112539

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Multi-view distillation based on multi-modal fusion for few-shot action recognition (CLIP-MDMF)

Fei Guo [a,*], YiKang Wang [a], Han Qi [a], Wenping Jin [a], Li Zhu [a], Jing Sun [b]

[a] School of Software Engineering, Xi'an Jiaotong University, China
[b] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China

ARTICLE INFO

Keywords:
Few-shot action recognition
CLIP
Multi-modal
Multi-view
Cross-transformer
Distillation

ABSTRACT

In recent years, the field of few-shot action recognition (FSAR) has garnered significant attention. Although many methods primarily rely on mono-modal data, there is a growing trend towards utilizing multi-modal data. However, existing FSAR methods often employ simplistic fusion techniques, such as concatenation or comparison, which may not fully leverage the potential of multi-modal information. We aim to explore multi-modal information from different views and subsequently perform multi-view fusion. Based on the textual and visual modality information extracted by the CLIP backbone, we propose an MDMF method that comprehensively utilizes these modalities at two levels. At the first level, we extract visual information from two views: Local Temporal Context and Global Temporal Context. Within each view, we fuse visual features with textual information through concatenation and Cross-Transformer operations. We then employ metric comparison to derive probability distributions for classification under the meta-learning paradigm for each view. At the second level, we fuse the probability distributions from both views to make the final decision. Concurrently, during training, for each query, we calculate the posterior distributions of the textual and visual modalities within each view using text information distance and visual information distance. Based on these distributions, we group query samples with higher view reliability. Subsequently, we enhance the representation of the less reliable view of specific samples through mutual distillation. By delving deep into multi-modal data through a multi-view approach, our few-shot action recognition model demonstrates the potential for achieving higher accuracy and enhanced robustness. Our code is available at the URL: https://github.com/cofly2014/CLIP-MDMF.

1. Introduction

Few-shot action recognition tackles two primary challenges: representing distinguishable spatiotemporal sequences with limited samples and establishing sequence comparisons between support and query instances. Recent approaches have delved into multi-modal solutions for these challenges, especially for combining textual and visual modalities.

There have been several works belonging to multi-modal for few-shot action recognition. CMN-J [1] is an early work that belongs to this category. On one hand, it introduces a label-independent repository to store correlations between unlabeled data and target samples. On the other hand, it uses multi-modal features that contain RGB and optical flow features. ARCMN [2] introduces the cross-modal: one modal is the video content, and another modal is the action label. AMeFu-Net [3] incorporates both the RGB feature and depth features. MORN [4] introduces CLIP [5] to few-shot action recognition and brings multi-modal. It uses the semantics of labels to enhance the prototype and uses TRX for metric comparison. CLIP-FSAR [6] is a work also related to CLIP. Firstly, it uses the average for all the video frames and uses text–visual comparison for CLIP adaptation. Then, in each episode, it concatenates the prompt embedding and the visual feature for constructing a multi-modal prototype. It only uses the labels for support, so perhaps it has not utilized labels efficiently. AMFAR [7] utilizes RGB and optical flow modalities alongside distillation techniques.

These multi-modal methods do not delve deeply into the visual modal, and the fusion methods used in the aforementioned approaches are relatively simplistic. Since the focus of video actions can vary, with some emphasizing global information and others highlighting local information, both global and local information are crucial for accurately determining the action category. By extracting and combining insights from both local and global context views, we can obtain more comprehensive and complementary information. In each context view, we fully exploit both textual and visual modalities, deeply fuse them, and use the fused features for classification. Our model pays particular attention

* Corresponding author.
E-mail addresses: co.fly@stu.xjut.edu.cn (F. Guo), funnyQ@stu.xjtu.edu.cn (Y. Wang), qihan19@stu.xjtu.edu.cn (H. Qi), jinwenping@stu.xjtu.edu.cn (W. Jin), zhuli@xjtu.edu.cn (L. Zhu), jing.sun1@siat.ac.cn (J. Sun).

https://doi.org/10.1016/j.knosys.2024.112539
Received 13 June 2024; Received in revised form 16 August 2024; Accepted 21 September 2024
Available online 25 September 2024

to features derived from multi-view, based on multi-modal fusion. The model operates on two levels: the first level enhances multi-modal supplementation between label textual and visual information, while the second level fuses multi-views and conducts mutual distillation between views of global and local temporal contexts. Our work contains three important points as follows: (1) Efficient label utilization: We aim to utilize labels not only for support samples but also for query samples, even within the meta-learning paradigm's absence of query labels during meta-training. (2) Enhanced multi-modal fusion: We seek to improve the fusion of label prompt embedding and visual embedding to more effectively utilize multi-modal information under each view. (3) Robust model construction: By further exploiting fused multi-modal information, we aim to reduce data distribution bias through multi-view classification probability fusion and multi-view distillation across different temporal contexts. **Note that the points (1) and (2) belong to the first level, and the point (3) belongs to the second level.**

From a technical perspective, we propose Multi-view Distillation based on Multi-modal Fusion, utilizing CLIP [5] as the backbone. We describe and address these points as follows: (1) The label prompt describing a specific category of videos has consistency for all videos in such a category. Therefore, it robustly represents videos and contributes to the stability and distinguishability of inter-class distribution. For CLIP-FSAR [6], in each episode, label prompt embedding and visual embedding are concatenated for support to counteract sample-specific distributions. However, for queries that lack labels, only visual embedding is used. This results in inconsistency in the amount of information between query and support. We propose a Probability Prompt Selector to solve this problem. In an N-way K-shot setting, the query category must match one of the support categories. We compare the visual embedding of the query with the prompt embedding of supports to obtain a set of matching scores and convert them into a probability distribution. Based on the probability, we select the prompt embedding for the query through uniform sampling. (2) We use two different Context Extractors to obtain visual features from two different views. A Local Temporal Context Extractor is used to encourage the extraction of local sequence information, while a Global Temporal Context Extractor is used to promote the extraction of global sequence information. For each view, we use the Cross-Transformer to fuse the prompt embedding, local (global) context, and regular visual features to obtain more distinguishable features. For the features of each view, we obtain two different metric comparison classification probabilities (one for textual-modal and the other for visual-modal) in each episode. (3) Then, we fuse the classification probabilities from multi-view as the final decision. Additionally, we perform knowledge mutual distillation between the two views to enforce consistent class prediction between global and local context views. Depending on the posterior distribution of textual-modal comparison and visual-modal comparison, the distillation direction for each query is different. Ultimately, our goal is to deeply explore multi-modal features to mitigate inter-class overlap and bias while enhancing model robustness through Multi-view information complementarity.

Our contributions are outlined as follows:

- (1) We first propose a framework for the Multi-view Distillation based on Multi-modal Fusion.
- (2) In the multi-modal prototype matching paradigm based on CLIP, we propose a new concept of Probability Prompt Selector to compensate for the information inconsistency between prototypes and queries to utilize labels efficiently.
- (3) We propose two different Context Extractors to get the different context features from two views. In each view, a Cross-Transformer is used to fuse the prompt embedding, visual feature, and visual context.
- (4) We use distance fusion and mutual distillation between Multi-view to enhance the performance further.
- (5) A large number of experiments on five benchmarks, including HMDB51, UCF101, Kinetics, SSV2-Full and SSv2-Small, demonstrate the rationality of our setting and the effectiveness of our proposed method. The results can be compared with the state-of-the-art methods.

2. Related work

In this section, we briefly review several related works. We first introduce the works of few-shot image classification in Section 2.1. Then, we review the works of few-shot action recognition in Section 2.2. Also, we review the CLIP and knowledge distillation in Sections 2.3 and 2.4.

2.1. Few-shot image classification

Few-shot image classification aims to identify objects of unseen categories using only a few labels and also a large number of samples under the seen categories. Research in this field is broadly divided into three categories: metric-based, optimization-based, and augmentation-based. Metric-based methods [8–13] extract the spatiotemporal features and use support-query matching rules to classify the query. The matching rules contain *cosine* similarity, Euclidean distance, and learnable distance based on neural networks. Optimization-based methods make the provided model well-initialized and easy to reach the optimal point, just as MAML [14–19] and related variants. Augmentation-based methods [20–25] make use of generative strategies that could produce lots of valuable data under conditions without enough data.

2.2. Few-shot action recognition

Unlike image classification, action recognition must consider the temporal dimension and the extraction of keyframes. Few-shot action recognition aims to solve the unrealistic problem of acquiring lots of labeled video samples. There are several important works in this field. CMN [26] adopts a memory network to store representations and classify action videos through matching and sorting. OTAM [27] computes a distance matrix of frames using the DTW [28] method and conducts strict matching. TRX [29] employs subsequence cross-attention to extract feature prototypes across different temporal scales, effectively mitigating temporal misalignment. STRM [30] enhances TRX features through preprocessing for feature enrichment. TSA-MLT [31] filters out irrelevant frames and makes efficient use of tuple alignment across different levels in TRX. MTFAN [32] introduces an end-to-end network by jointly exploring task-specific motion modulation and multi-level temporal fragment alignment. MoLo [33] devises a motion-augmented long-short contrastive learning method to jointly model global contextual information and motion dynamics. ProtoGAN [34, 35] use GAN [36] to synthesize additional data for novel classes. CLIP-FSAR [6] harnesses CLIP's strong generalization ability, trained on extensive datasets, utilizing encoders to encode text and images, followed by Transformer enhancement for comparison. AMFAR [7] employs bidirectional distillation to capture differentiated task-specific knowledge from reliable modalities, enhancing the representation of unreliable modalities. Our work, like CLIP-FSAR, is also based on CLIP. In AMFAR, the important modality of optical flow data is involved, but such data is often difficult to obtain. CLIP-FSAR effectively applies the CLIP model and integrates textual and visual modalities. However, further exploration is needed to better utilize the information from both textual and visual modalities.

2.3. CLIP model

The CLIP model [5] is a pre-trained neural network model released by OpenAI in early 2021 for matching images and texts. It has been a classic in multi-modal research in recent years. The model directly uses a large amount of text-image pairs from the Internet for pre-training and has achieved the best performance in many tasks. Using these pairs for training can not only solve the high-cost problem of obtaining labeled data but also make it easier to obtain a model with solid generalization ability depending on the amount of data from the network, which is relatively large, and the data is pretty different. The idea of the CLIP model is to improve the model's performance by learning the matching relationship between image and text. Precisely, the CLIP model consists of two main branches: an encoder with CNN-RN [37] or the VIT [38] for processing images and a Transformer model [39] for processing text. Both branches are trained to map the input feature into the same embedding space and force image-text pairs to be close in the embedding space. Except for the image-text, some other tasks also use the CLIP model for the multi-modal comparison, and how to give a perfect method for the domain adaption is essential.

2.4. Knowledge distillation

Knowledge distillation is the classic model compression method [40–42], with the core idea of guiding lightweight student models to mimic better-performing and more structurally complex teacher models. Optimization strategies, such as mutual learning and self-learning through neural networks and data resources, such as unlabeled and cross-modal, significantly enhance model performance. Knowledge Amalgamation [43] is the migration of multiple tasks into a single student model to make it capable of handling multiple tasks. Mutual distillation [44] involves student models learning from each other to improve performance without relying on a robust teacher network, thereby avoiding the dependence on large-scale teacher models. ER-SCMT [45] proposed cross-modal affective recognition with the distillation of data features from different modalities. AMFAR [7] is related to the mutual distillation of multi-modal. Our work is based on the Multi-view distillation.

3. Methodology

In this section, we first describe the formulation of few-shot action recognition in Section 3.1. Then, we present the framework of the proposed CLIP-MDMF in Section 3.2. In Sections 3.3, 3.4 and 3.5, based on the three points proposed in the Introduction, we respectively introduce the Probability Prompt Selector for efficient label utilization, Multi-modal fusion under Multi-view for enhanced fusion, Multiple-view fusion and Mutual Distillation for robust model construction. At last, in Section 3.6, We briefly described the inference method.

3.1. Problem setting

In the field of few-shot action recognition, the video datasets are split into D_{train}, D_{test}, D_{val}, all the split datasets should be disjoint, which means there are no overlapping classes between each split dataset. In the D_{train}, it contains abundant labeled data for each action class, while there are only a few labeled samples in D_{test}, and D_{val} is used for model evaluation during the training episode. No matter D_{train}, D_{test}, or D_{val}, they all follow a standard episode rule. The episode, also called a task, occurs during the training, testing, or validation. In each episode, N classes with K samples in D_{train}, D_{test}, or D_{val} are sampled as "support set". The samples from the rest videos of each split DB are sampled as "query set", just as P samples are selected from N classes to construct the "query set". The goal of few-shot action recognition is to train a model using D_{train}, which can be generalized well to the novel classes in D_{test} only using $N \times K$ samples in the support set D_{test}. Let $q = (q_1, q_2, \ldots, q_m)$ represents a query video with m uniformly sampled frames. We use $C = \{c_1, \ldots, c_N\}$ to represent the class set, and we aim to classify a query video q into one of the classes $c_i \in C$. In our work, the support set is defined as S, and the query set is defined as Q. For the class c, the support set s_c can be expressed as $s_c = \{s_c^1, \ldots, s_c^k \cdots, s_c^K\}, 1 \le k \le K$, and $s_c^k = (s_c^{k,1}, \ldots, s_c^{k,m}), 1 \le k \le K$, m is the frame number.

3.2. Overview

The framework of our model is shown in Fig. 1. **Firstly**, the visual encoder of CLIP is used to get the visual embedding of support and query. The text encoder of CLIP is used to get the label prompt embedding of the support categories in each episode. Also, a Probability Prompt Selector (PPS) is proposed to generate a probability prompt for each query video. The prompt embeddings introduce the stable feature that does not change with sample distribution and ensure the consistency of information between the representation of support and query. In this part, the inputs are video frames (both the support and query videos) and video labels (only the support videos). The outputs are visual embedding of support and query, label prompt embedding of the support categories, and probability prompt of the query video. **Secondly**, we introduce the Multi-view structure, where each view fuses the multi-modal related to the label feature and visual feature. The Multi-view structure is as follows: (1) Local Temporal Context Extractor (LTCE). Using several Conv1d operations in the temporal dimension, each frame could contain the context information of adjacent frames. (2) Global Temporal Context Extractor (GTCE). Using the TCN [46] network in the temporal dimension, each frame could get the global sequence context. (3) Multi-modal Fusion Encoder (MMFE). The core of MMFE is a Cross-Transformer, which is introduced for each view to extract multi-modal features related to the label prompt and visual. This module concatenates each video context from the LTCE (or GTCE) with prompt embedding as the $Query$ and concatenates the features from the CLIP visual encoder with prompt embedding as Key and $Value$. The fused features are then obtained through the Cross-Transformer. In this part, the inputs are visual embedding of support and query, label prompt embedding of the support categories, and probability prompt of the query video. The outputs are the fused features of the support and query in both the global context view and the local context view. **Thirdly**, fusion and distillation of the two views enable the model to register the multi-modal features from global and local temporal contexts, thereby enabling the model to learn more general features. In this part, the inputs are the fused support and query features of two views. The outputs are the loss resulting from the fusion of two views at the decision level, as well as the loss from the mutual distillation of the two views. These losses are used to optimize the network.

3.3. Probability Prompt Selector (PPS)

The existing CLIP-based few-shot action recognition works, as CLIP-FSAR [6] and MORN [4] incorporate text information while constructing support prototype. However, the query still maintains the mono-modality of visual. Intuitively, the amount of information between the support prototype and the query is inconsistent. Because the label prompt embedding under the same category has consistent features across all videos, it does not change with different video instances and is robust to the probability distribution of representations. So, it is essential to introduce label prompt embedding into the query. In the paradigm of meta-learning, although there is no query label beforehand, it must belong to the label set of the support in each episode. Given a query q, we assume the representation from the visual encoder is f_q, and label prompt embedding of support class c from the text encoder is $token_c$, then we calculate the $cosine$ similarity between f_q and $token_c$:

$$sim_c^q = \frac{< f_q, token_c >}{|f_q| \cdot |token_c|} \tag{1}$$

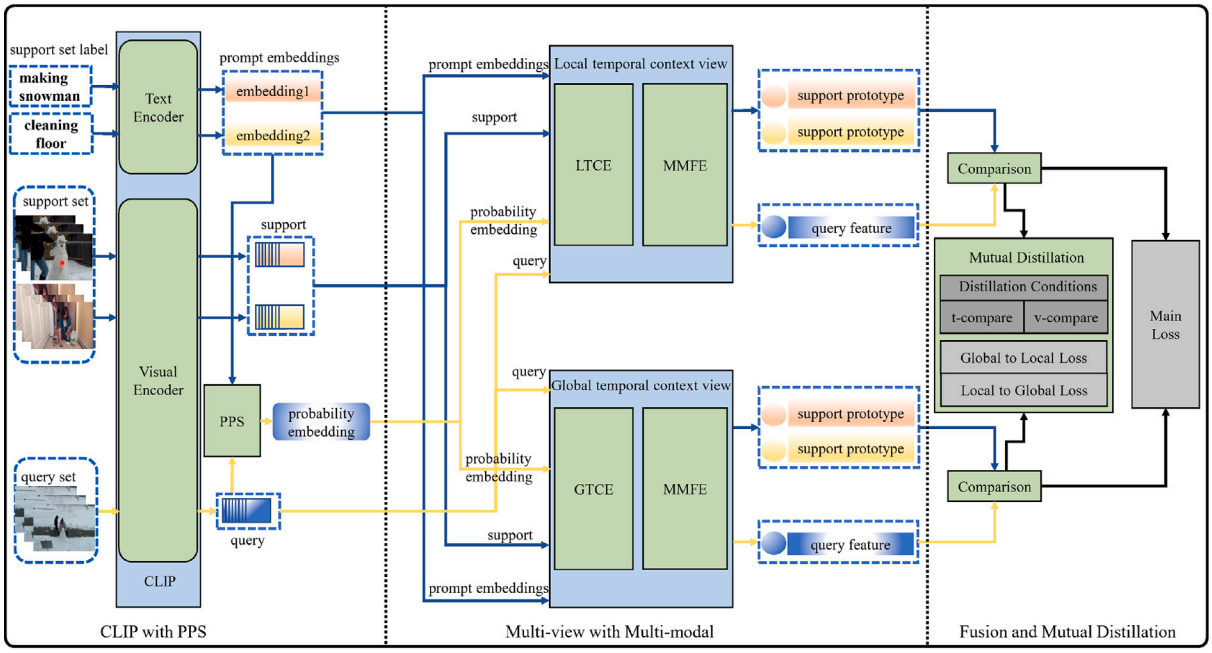


Fig. 1. The framework. The green regions are the modules or some operations. There are 6 modules in our model. (1) CLIP model. (2) Probability Prompt Selector (PPS). (3) Local Temporal Context Extractor (LTCE). (4) Global Temporal Context Extractor (GTCE). (5) Multi-modal Fusion Encoder (MMFE). (6) Multiple-view Mutual Distillation(MVMD). There are some other operations: (a) Comparison. (b) Main loss. The blue arrows are the data flow of support. The yellow arrows represent the data flow of the query, including the visual and label. The black arrows represent the decision data flow of each view. The area with a light blue background is the modules that are related to the training. The MMFE in the Local temporal context view and Global temporal context view have the same structure but different parameters.

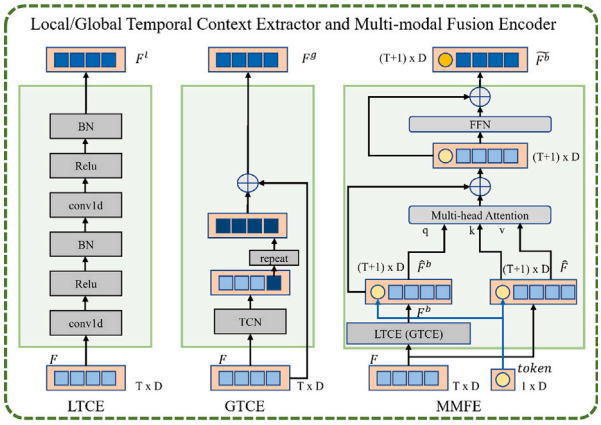


Fig. 2. The left part is the LTCE. It contains several Conv1d operations. According to Conv1d, the features focus on the local temporal context. The middle part is the GTCE, and the core part is TCN. We add the last frame of the output to each input frame. The features focus on the global temporal context. The right part is the MMFE, which is based on a Cross-Transformer. Features from the LTCE (GTCE) are concatenated with prompt embedding. The same operation is for the features that are directly from the CLIP visual encoder. Then, the LTCE (or GTCE) concatenated features are used as the queries, and the CLIP visual concatenated features are used as the keys and values.

Then, we use the Softmax with the temperature coefficient t to transfer the similarity value into a probability distribution.

$$prob_c^q = \frac{exp(sim_c^q/t)}{\sum_{c' \in C} exp(sim_{c'}^q/t)} \tag{2}$$

where $C = \{c_1, \ldots, c_N\}$ is the category set in each episode. According to the probability distribution, we use uniform sampling to sample the prompt embedding in the label set of support for the query q. Through the PPS, we obtain the embedding for the query.

However, the matching of video-text is still not enough for video. We finally need to make use of the comparison between query and prototype. Extracting more information from frame sequences and integrating the label prompt information into visual information is essential.

3.4. Multi-modal fusion under multi-view

Our work considers the fusion of multi-modal features from two temporal context views. The first one is the local temporal context view (LTC view), which contains LTCE and MMFE. The second is the global temporal context view (GTC view), which contains GTCE and MMFE.

3.4.1. Local Temporal Context Extractor (LTCE)

Fig. 2 (Left) depicts the illustration of LTCE. This Extractor includes a series of operations such as Conv1d, Relu, BN, etc. After these operations, the features pay more attention to the information of adjacent frames in front or after, thus getting the local temporal context of the features. For each video, we select 8 frames as the full sequence and use a Convolution kernel with a size equal to 3 to get the local temporal context. Given the features $F = [f^1, f^2 ..., f^T] \in R^{T \times D}$ (D is the dimension of the frame feature, and T is the frame number) from the CLIP visual encoder, they are processed by the LTCE. Eq. (3) shows the operations in the LTCE.

$$\begin{aligned} &F^1 = F * W_1, F^2 = RELU(F^1), F^3 = BN(F^2) \\ &F^4 = F^3 * W_2, F^5 = RELU(F^4), F^l = BN(F^5) \end{aligned} \tag{3}$$

where W_1 and W_2 are Convolution kernels. $F^i, i \in \{1, \ldots, 5\}$ is the temporary variable. F^l is the local temporal context.

3.4.2. Global Temporal Context Extractor (GTCE)

Fig. 2 (Middle) depicts the illustration of GTCE. Given the features $F = [f^1, f^2 ..., f^T] \in R^{T \times D}$ (D is the dimension of the frame feature, and T is the frame number) from the CLIP visual encoder. We use TCN [46] to extract the global temporal features. To be specific, we use a TCN with three layers to get the temporal features of frames. Because the TCN uses Dilated Convolution and Causal Convolution, the dilated rate grows exponentially by 2. When $T = 8$, the last frame of the output

feature can capture the temporal context for the full sequence. We copy the last frame of the output for T copies and add the input as the global temporal context. See Eq. (4).

$$F^g = Repeat(TCN(F)[-1]) + F \tag{4}$$

3.4.3. Multi-modal Fusion Encoder (MMFE)

In Sections 3.4.1 and 3.4.2, we get different temporal contexts from two views. Now, we study how to fuse the original sequence with the local temporal context(or global temporal context) and the corresponding prompt embedding. The fusing operations for the local and global temporal views are similar. Fig. 2 (Right) shows the Multi-modal Fusion Encoder using a Cross-Transformer.

Multi-modal Feature Concatenation. Given the features $F = [f^1, f^2, \dots, f^T]$ from the CLIP visual encoder, the temporal context feature $F^b = [f^{b,1}, f^{b,2}, \dots, f^{b,T}]$ from the LTCE (or GTCE), and the corresponding *token* (prompt embedding) from the CLIP text encoder or the PPS, we concatenate the prompt embedding from the text encoder with the support visual features and concatenate the prompt embedding from the text encoder with the local (or global) temporal context of support, respectively. We also concatenate the probability prompt embedding from PPS with query visual features and concatenate the probability prompt embedding with the local (or global) temporal context of the query. See Eqs. (5) and (6)

$$\hat{F} = Concat(token, F) \tag{5}$$

$$\hat{F}^b = Concat(token, F^b) \tag{6}$$

where b means the view, which can be g or l to represent the global or local temporal context (Refer to Sections 3.4.1 and 3.4.2). F and F^b represent both the support and query.

Multi-modal Feature Extraction. We use $\hat{F}^b$ as the *Query*, and $\hat{F}$ as the *Key* and *Value*, and send them into a Cross-Transformer.

$$\widetilde{F^b} = \mathbb{T}(\hat{F}^b + f_{pos}, \hat{F} + f_{pos}, \hat{F} + f_{pos}) \tag{7}$$

where $\mathbb{T}$ is the Transformer that contains the Multi-head Attention and FFN. $f_{pos} \in R^{(T+1)\times D}$ means the position embeddings to encode the position. $\widetilde{F^b} \in R^{(T+1)\times D}$.

3.5. Multiple-view fusion and Mutual Distillation (MVMD)

3.5.1. Distance fusion of different temporal context view

For each view through the MMFE, in each episode, we assume the sample feature under support class c as $\widetilde{F}^b_{s^k_c}$ where $k \in \{1, \dots, K\}$ and the feature of a query as $\widetilde{F}^b_q$. Here, b can be g or l as the global or local temporal context view. Using the average aggregation for the support features, the prototype is calculated as follows:

$$U^b_c = \frac{1}{K}\sum_{k=1}^{K} \widetilde{F}^b_{s^k_c} \tag{8}$$

where the s^k_c is defined in Section 3.1, it means the kth support sample in the class c.

We only use the visual frames (not including the first item of the query and the support prototype) from the MMFE to calculate the distance. We calculate the distance of the global context view as the $dis(\widetilde{F^g_q}, U^g_c)$ and the distance of the local context view as $dis(\widetilde{F^l_q}, U^l_c)$.

$$dis_{total}(q, s_c) = \alpha_1 dis(\widetilde{F^g_q}, U^g_c) + \alpha_2 dis(\widetilde{F^l_q}, U^l_c) \tag{9}$$

where α_1 and α_2 are the hyper-parameters. s_c is defined in Section 3.1, which means the support samples of class c.

Using Softmax for $dis_{total}(s_c, q)$, we can get the classification probability, see Eq. (16). We assume L_{main} is the Cross-Entropy Loss between the probability and the ground truth.

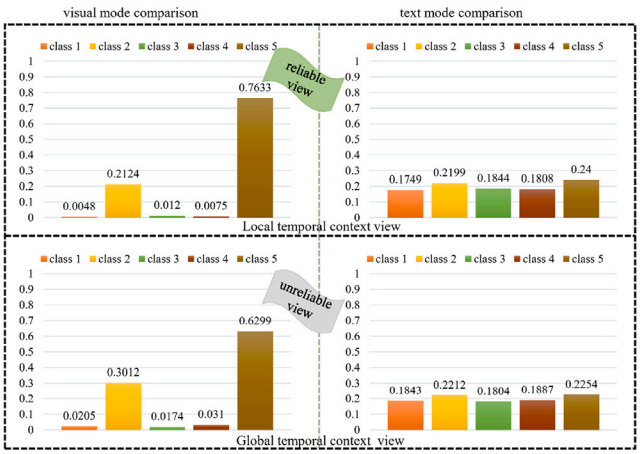


Fig. 3. The distillation condition. The upper is the local temporal context view for visual comparison and text comparison. The lower is the global temporal context view for visual comparison and text comparison. We can see the $0.7633 > 0.6299$ and $0.24 > 0.2254$, so the local temporal context view is reliable.

3.5.2. Distillation conditions

Inspired by the work [7], we select queries with significant differences in reliability between the two views, where the more reliable view is regarded as the primary view. We define the reliable view of each query as reflecting more discriminative features of specific tasks, so it deserves more attention in few-shot learning. For a query, the reliable view may vary across different tasks, as the contribution of a specific view largely depends on the context information of the query and support in each episode. In Fig. 1, we can see the output of the Local (Global) Temporal Context view contains two modalities: visual modal and textual modal. Our distillation conditions are based on the view-specific posterior distribution for both the visual embedding comparison and the label prompt embedding comparison. Given a query q and its probability prompt label $lable_q$, the view-specific posterior distributions of two modalities are as follows:

$$\mathcal{P}^b(y^{visual}_q = c|q, label_q) = \frac{exp(-dis(\widetilde{F^b_q}, U^b_c))}{\sum_{c' \in C} exp(-dis(\widetilde{F^b_q}, U^b_{c'}))} \tag{10}$$

$$\mathcal{P}^b(y^{text}_q = c|q, label_q) = \frac{exp(cosin(token^b_q, Token^b_c))}{\sum_{c' \in C} exp(cosin(token^b_q, Token^b_{c'}))} \tag{11}$$

Similar to Section 3.5.1, b can be g or l as the global or local temporal context view. The *visual* means the visual modal and *text* means the textual modal. We define $token^b_q$ as the first item of $\widetilde{F^b_q}$ and $Token^b_c$ as the first item of U^b_c where $c \in [c_1, \dots, c_N]$. $Token^b_c$ is prompt embedding for the prototype U^b_c which is belongs to category c and the view b.

We define the distillation discriminant score as the maximum element of the view-specific posterior distribution.

$$\hat{c}^b_q = \max_k \mathcal{P}^b(y^{visual}_q = k|q, label_q) \tag{12}$$

$$\tilde{c}^b_q = \max_k \mathcal{P}^b(y^{text}_q = k|q, label_q) \tag{13}$$

where for the query q, $\hat{c}^b_q$ is the discriminant score of visual modal, and $\tilde{c}^b_q$ is discriminant score of textual modal.

For the query, if a specific view achieves a higher discriminant score in both the visual and textual modal than another view, this specific view is reliable for expressing discriminative action features in each episode. On the contrary, if the discriminant score is lower in both modalities, the specific view is unreliable in identifying actions. If the specific view achieves a higher discriminant score only in one mode, we do not define the reliable or unreliable view, and we do not use the distillation for the query. As shown in the example of Fig. 3, the local temporal context view is reliable. We define the set of global view reliable samples as Ω^g and the set of local view reliable samples as Ω^l.

3.5.3. Mutual distillation

Depending on the $\hat{c}_q^b$ and $\tilde{c}_q^b$ where $b \in \{g, l\}$, KL divergence is used for the mutual distillation. For each query, the reliable view acts as the teacher, and the unreliable view acts as the student. Mutual distillation Losses are as follows:

$$L_{l\to g} = \frac{1}{\sum_{q\in\Omega^g}(\hat{c}_q^g + \tilde{c}_q^g)} \sum_{q\in\Omega^g} (\hat{c}_q^g + \tilde{c}_q^g) D_{KL}(P_q^g, P_q^l)$$
$$L_{g\to l} = \frac{1}{\sum_{q\in\Omega^l}(\hat{c}_q^l + \tilde{c}_q^l)} \sum_{q\in\Omega^l} (\hat{c}_q^l + \tilde{c}_i^l) D_{KL}(P_q^l, P_q^g) \tag{14}$$

where P_q^l and P_q^g are calculated as Eq. (10) for classification distribution of visual modal.

The final loss can be denoted as:

$$L = L_{main} + \lambda(L_{l\to g} + L_{g\to l}) \tag{15}$$

where λ is the hyper-parameter.

3.6. Inference

In the meta-testing stage, we fuse the distance of the local temporal context view and the distance of the global temporal context view for inference. Refer to Eq. (9). Using Softmax for the dis_{total} of all the support prototypes in each episode, we can get the classification probability for inference.

$$P(y = c|q) = \frac{exp(dis_{total}(s_c, q))}{\sum_{c'\in C} exp(dis_{total}(s_{c'}, q))} \tag{16}$$

4. Experiments

In this section, we first introduce four challenging datasets and the split methods. Second, we describe the implementation details of our work. After that, we show the comparison experiment for our proposed CLIP-MDMF and some other state-of-the-art methods. Then, we do some ablation studies to demonstrate the effectiveness of the key components of our method. Also, there are some other experiments related to distribution comparison, accuracy comparison, and attention visualization comparison to further prove the progressiveness of our method.

4.1. Datasets

In our experiments, we utilize the datasets including UCF101 [47], HMDB51 [48], Kinetics-400 [49], SSv2-Full and SSv2-Small [50]. The dataset settings for SSv2-Small and Kinetics-400 are as follows: the split methods for HMDB51 and UCF101 adhere to the protocol established by ARN [51]. Specifically, UCF101 is divided into 70 classes for training, 10 for validation, and 21 for testing, with 9154, 1421, and 2745 videos allocated to the train, validation, and test sets. HMDB51 comprises 31 classes for training, 10 for validation, and 10 for testing, containing 4280, 1194, and 1292 videos for the train, validation, and test sets. For the Kinetics-400 and SSv2-Small settings, we employ the split methods proposed by CMN [26] and CMN-J [1]. These methods involve a random selection process that creates a mini-dataset with 100 classes, which includes 64 classes for training, 12 for validation, and 24 for testing, with each class containing 100 samples. We also follow the OTAM [27] split method for configuring the SSv2-Full dataset, which contains 77,500 videos for training, 1926 for validation, and 2854 for testing across the train, val, and test sets. The class division for SSv2-Full mirrors that of SSv2-Small but with a larger number of samples per class.

4.2. Implementation details

Data augmentation: in the training stage, we flip each frame horizontally and randomly crop the center region 224×224. Backbone: we use both the ResNet50 [37] and VIT-B/16 [38] of CLIP as the visual encoder. Optimizer: we use the Adam [52]. Learning rate: the learning rate is 0.00001. Video frames: we follow the previous work TSN [53] for the video frame. Eight frames are sparsely and uniformly sampled from each video. Training stage: we do not freeze CLIP; instead, we perform joint optimization with other modules based on the pre-trained CLIP model. We average gradients and backpropagate once every 16 iterations. The modules that require parameter updates include CLIP, LTCE, GTCE, and MMFE. Testing stage: we run 10,000 episodes, and our experiment's average accuracy is reported. For the augmentation, we use only the center crop to augment the video. We use the OTAM [27] as the comparison method for distance.

4.3. Comparison with the-state-of-the-art works

We compare our model with the state-of-the-art methods, and our baseline is the CLIP-FSAR. Firstly, we use CLIP-RN50 as the backbone. For the UCF101 and HMDB51, see in the Table 1: in the 5-shot setting, our CLIP-MDMF is significantly superior to CLIP-FSAR with 1.8% and 2.3%. In the 3-shot setting, our CLIP-MDMF is superior to CLIP-FSAR from 95.4% to 98.0% and from 78.3% to 79.4%. Also, for UCF101, in the 1-shot setting, our CLIP-MDMF is superior to CLIP-FSAR from 92.4% to 94.3%. For the Kinetics, SSv2-Full and SSv2-Small, see in the Table 2: in the 5-shot setting, our CLIP-MDMF is still superior to CLIP-FSAR with 1.5%, 5.2% and 4.6%. In the 3-shot setting, our CLIP-MDMF is superior to CLIP-FSAR from 90.8% to 92.0%, from 60.7% to 63.9% and from 54.0% to 56.3%. For the Kinetics, in 1-shot setting, the accuracy of our CLIP-MDMF is equal to CLIP-FSAR. For the SSV2-Small, in the 1-shot setting, our CLIP-MDMF is superior to CLIP-FSAR from 52.1% to 52.3%. With CLIP-RN50, compared to AMFAR, the accuracies under HDMB51 and SSV-Full are lower. However, AMFAR introduces optical flow data, which is sometimes difficult to obtain in the production environment. We also give the results based on CLIP-VIT. In SSv2-Full, the AMFAR is better than ours because SSv2 is a time-sensitive dataset. The increased optical flow information has made a significant contribution. For other datasets, our model also gets lots of state-of-the-art results.

According to Tables 1 and 2, we can conclude that: (1) Our model is better than most of the current methods. (2) Our model is better than CLIP-FSAR. (3) Our model with CLIP-RN50 is competitive with AMFAR. Except for SSv-Full, our model with CLIP-VIT is better than AMFAR even though optical flow data are used in AMFAR.

The videos in the UCF dataset feature a variety of backgrounds, lighting conditions, and video qualities. SSV2 videos are characterized by their temporal sensitivity. The Kinetics dataset excels in capturing spatially sensitive sequences. In the HMDB51 dataset, the speeds and rhythms of action execution vary. As a result, both global and local perspectives can extract significantly different features from these datasets. In summary, as shown in Tables 1 and 2, based on CLIP-RN50, our model adapts well to action recognition datasets with different characteristics by integrating decisions from different views. Referring to the ablation study, we find that our use of the Probability Prompt Selector to choose labels for query samples, along with the Local Temporal Context Extractor and Global Temporal Context Extractor for extracting video information from different perspectives, enhances performance through view result distillation and fusion at the decision level. This contributes to the improved performance of our model. However, the relative improvement of our model with the CLIP-VIT backbone is minimal. This is because CLIP-VIT possesses a more powerful feature extraction capability, which may overshadow the advantages of our model.

Table 1
Comparison with the state-of-the-art few-shot action recognition methods on the UCF101, HMDB51 datasets. We report the results on 5-way 5-shot, 3-shot, and 1-shot settings, "–" means the results are not available in published works, ↑ means better than baseline CLIP-FSAR, and ⋄ means our implementation (see [3,6,7,9,27,29,30,32,33,54–59]).

Method	Reference	Backbone	HMDB51			UCF101		
			1-shot	3-shot	5-shot	1-shot	3-shot	5-shot
ProtoNet [9]	NeurIPS'2017	ResNet-50	54.2	–	68.4	70.4	–	89.6
OTAM [27]	CVPR'2020	ResNet-50	54.6	–	66.1	79.9	–	88.9
TRX [29]	CVPR'2021	ResNet-50	52.9°	–	75.6	77.3°	–	96.1
STRM [30]	CVPR'2022	ResNet-50	54.1°	–	77.3	79.2°	–	96.9
HyRSM [54]	CVPR'2022	ResNet-50	60.3	71.7	76.0	83.9	93.0	94.7
MTFAN [32]	CVPR'2022	ResNet-50	59.0	–	74.6	84.8	–	95.1
TA2N [55]	AAAI'2022	ResNet-50	59.7	–	73.9	81.9	–	95.1
HCL [56]	ECCV'2022	ResNet-50	59.1	–	76.3	82.6	–	94.6
MPRE [57]	TCSVT'2022	ResNet-50	57.3	–	76.8	82.0	–	96.4
TADRNet [58]	TCSVT'2023	ResNet-50	64.3	74.6	78.2	86.7	94.3	96.4
MoLo [33]	CVPR'2023	ResNet-50	60.8	72.0	77.4	86.0	93.5	95.5
AMeFu-Net [3]	ACMMM'2020	ResNet-50	60.2	–	75.5	85.1	–	95.5
SRPN(2021) [59]	ACMMM'2021	ResNet-50	61.6	72.5	76.2	86.5	93.8	95.8
AMFAR [7]	CVPR'2023	ResNet-50	73.9	–	87.8	91.2	–	99.0
CLIP-FSAR [6]	IJCV'2023	CLIP-RN50	69.4	78.3	80.7	92.4	95.4	97.0
CLIP-FSAR [6]	IJCV'2023	CLIP-VIT-B	77.1	84.1	87.7	97.0	98.6	99.1
CLIP-MDMF		CLIP-RN50	66.8	79.4 ↑	83.0↑	94.3 ↑	98.0↑	98.8↑
CLIP-MDMF		CLIP-VIT-B	77.1	84.1 ↑	88.0↑	97.0	98.1	99.3↑

Table 2
Comparison with the state-of-the-art few-shot action recognition methods on the Kinetics, SSv2-Small, and SSv2-Full datasets. We report the results on 5-way 5-shot, 3-shot, and 1-shot settings, "–" means the results are not available in published works, ↑ means better than baseline CLIP-FSAR, and ⋄ means our implementation (the data in parentheses represents the published data) (see [1,3,6,7,9,27,29,30,32,33,54–60]).

Method	Reference	Backbone	Kinetics			SSv2-Full			SSv2-Small		
			1-shot	3-shot	5-shot	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot
ProtoNet [9]	NeurIPS'2017	ResNet-50	65.4	–	77.9	–		–	33.6	–	43.0
Matching Net [60]	NeurIPS'2016	ResNet-50	53.3	–	74.6	–	–	–	34.4	–	43.8
OTAM [27]	CVPR'2020	ResNet-50	73.0	–	85.5	42.8	–	52.3	38.9°	–	48.1°
TRX [29]	CVPR'2021	ResNet-50	63.4°(63.6)	–	85.1°(85.9)	42.0	–	63.0°(64.6)	36.0	–	56.3°(59.4)
STRM [30]	CVPR'2022	ResNet-50	65.3°	–	85.9°(86.7)	42.9°	–	64.8°(68.1)		–	
HyRSM [54]	CVPR'2022	ResNet-50	73.7	83.5	86.1	54.3	65.1	69.0	40.6	52.3	56.1
MTFAN [32]	CVPR'2022	ResNet-50	74.6	–	87.4	45.7	–	60.4	–	–	–
TA2N [55]	AAAI'2022	ResNet-50	72.8	–	85.8	47.6	–	61.0	–	–	–
HCL [56]	ECCV'2022	ResNet-50	73.7	–	85.8	47.3	–	64.9	38.7	–	55.4
MPRE [57]	TCSVT'2022	ResNet-50	70.2	–	85.3	42.1	–	58.6	–	–	–
TADRNet [58]	TCSVT'2023	ResNet-50	75.6	84.8	87.4	43.0	–	61.1	–	–	–
MoLo [33]	CVPR'2023	ResNet-50	74.0	83.7	85.6	56.6	67.0	70.6	42.7	52.9	56.4
AMeFu-Net [3]	ACMMM'2020	ResNet-50	74.1	–	86.8	–	–	–	–	–	–
CMN++ [1]	TRAMI'2020	ResNet-50	60.5	75.6	78.9	36.2	44.6	44.8	–	–	–
SRPN [59]	ACMMM'2021	ResNet-50	75.2	84.7	87.1	–	–	–	–	–	–
AMFAR [7]	CVPR'2023	ResNet-50	80.1	–	92.6	61.7	–	79.5	–	–	–
CLIP-FSAR [6]	IJCV'2023	CLIP-RN50	90.1	90.8	91.6°(92.0)	58.7	60.7	62.9°(62.8)	52.1	54.0	55.3°(55.8)
CLIP-FSAR [6]	IJCV'2023	CLIP-VIT-B	94.8	95.0	95.4	62.1	68.3	72.1	54.6	59.4	61.8
CLIP-MDMF		CLIP-RN50	90.1	92.0 ↑	93.5↑	56.9	63.9 ↑	68.0 ↑	52.3↑	56.3 ↑	60.4 ↑
CLIP-MDMF		CLIP-VIT-B	95.0 ↑	95.1 ↑	96.2 ↑	60.1	68.9 ↑	72.7 ↑	54.0	59.8↑	62.8 ↑

4.4. Ablation study

To prove all the parts of our model are effective, we have designed several experiments. Firstly, we design an experiment to verify Single-view and PPS. Secondly, we do some experiments to verify the multi-view fusion of bidirectional distillation under both distillation conditions and PPS. Thirdly, an experiment is designed to determine the distillation condition and direction. Our ablation study is based on CLIP-RN50.

4.4.1. Probability prompt selector and single-view ablation

To demonstrate the effectiveness of LTCE (GTCE) in combination with MMFE and PPS, we conducted experiments comparing the views of GTC, LTC, and NTC. The baseline is NTC (None Temporal Context), which only employs a Transformer using the same *Query*, *Key*, and *Value* without processing by the Temporal Context Extractor. As shown in Table 3, regardless of whether LTC, GTC, or NTC view is utilized, in the majority of settings, the accuracy with PPS surpasses that without it. This suggests that the probability prompt for the query can indeed complement class consistency information. Furthermore, regardless of

Table 3
Single-view comparison for the temporal context view and PPS. GTC means Global Temporal Context, and LTC means Local Temporal Context. NTC means None Temporal Context is used, and it just employs a Transformer using the same *Query*, *Key*, and *Value* without processing by the Temporal Context Extractor. We report the results on the 5-way, 5-shot setting.

Single-view	Single-view			PPS	Kinetics	SSv2-Small	HMDB51
Serial No.	GTC	LTC	NTC		5-shot	5-shot	5-shot
1	✓	✗	✗	✗	92.0	58.1	82.4
2	✓	✗	✗	✓	93.2	58.6	82.4
3	✗	✓	✗	✗	91.9	56.2	82.1
4	✗	✓	✗	✓	92.9	57.2	82.3
5	✗	✗	✓	✗	91.4	56.0	81.7
6	✗	✗	✓	✓	92.0	56.5	82.2

Table 4
Multi-view distillation and PPS. We report the results on the 5-way, 5-shot setting.

Multi-view fusion Serial No.	Distillation	PPS	Kinetics 5-shot	SSv2-Small 5-shot	HMDB51 5-shot
1	✓	✗	92.3	57.1	82.4
2	✓	✓	93.5	60.4	83.0
3	✗	✗	92.2	56.6	82.1
4	✗	✓	93.1	59.1	82.5

the utilization of PPS, both the GTC view and LTC view exhibit higher accuracy compared to the NTC view, indicating their effectiveness.

4.4.2. Ablation for multi-view fusion with distillation and PPS

In Table 4, using distillation and PPS achieves the highest accuracy. The model with distillation is better than without it. The same result is for PPS. When the items are without PPS, the Multi-view distillation effect is limited. For Kinetics, the accuracy is from 92.2% to 92.3% (only 0.1% improvement), for SSv2-Small from 56.6% to 57.1% (0.5% improvement), and for HMDB51 from 82.1% to 82.4% (0.3% improvement). When we add PPS, the distillation effect increases. For Kinetics, the accuracy is from 93.1% to 93.5% (0.4% improvement), for SSv2-Small, from 59.1% to 60.4% (1.3% improvement), and for HMDB51 from 82.5% to 83.0% (0.5% improvement). For the items in Table 4 compared with the according items in Table 3, we wish the accuracies of the Multi-view always to be higher than the Single-view. But for SSv2-Small, we can see the accuracy of $No.3$ item without PPS and distillation (56.6%) and the accuracy of $No.1$ item only without PPS (57.1%) in Multi-view Table 4 are lower than the accuracy of $No.1$ item with GTCE but without PPS (58.1%) in the Single-view Table 3 . For the HMDB51, in the Multi-view Table 4, the accuracy of $No.1$ item only without PPS (82.4%) is equal to the accuracy of $No.1$ item with GTCE but without PPS (82.4%) in the Single-view Table 3. The reason for this is that we merely fuse the distances of the Multi-view, as seen in Eq. (9), without imposing any additional constraints. Therefore, the fusion distance or distillation might not consistently enhance discriminative capability without PPS.

4.4.3. Multi-view distillation condition and direction

In Table 5, v-compare signifies visual comparison, while t-compare denotes token comparison. Up and down indicate LTC view and GTC view, respectively. Leveraging both t-compare and v-compare methods, along with bidirectional mutual distillation, our model achieves the highest accuracy across Kinetics, SSv2-Small, and HMDB51 datasets. Combining both distillation conditions results in superior accuracy compared to using just one. Overall, bidirectional distillation outperforms unidirectional distillation in accuracy. Given the video samples, when the distillation condition is used, the reliable branch has the higher classification confidence (posterior probability). Bidirectional distillation involves transferring knowledge from the reliable branch of specific samples to the less reliable branch. Since the reliable branch varies across different samples, the direction of knowledge transfer is also different. Mutual distillation enables both branches to better adapt to the dataset. In contrast, unidirectional distillation only considers samples with higher classification confidence in one branch and transfers this knowledge to another. In Table 5, we observe that the performance of bidirectional distillation using t-compare is slightly lower than that of single-direction distillation, while the bidirectional distillation of v-compare performs better. This discrepancy can be attributed to the significant information differences in the visual components between LTCE and GTCE, which are extracted using Conv1d and TCN. Conversely, the differences in semantic distribution are relatively minor, as they are primarily shaped through processing by MMFE in Section 3.4.3. Consequently, relying solely on bidirectional distillation based on t-compare proves less effective in training a superior model.

Table 5
Distill condition and direction study.

Distillation condition t-compare	Distillation condition v-compare	Distillation direction	Kinetics 5-shot	SSv2-Small 5-shot	HMDB51 5-shot
✓	✓	bidirectional	93.5	60.4	83.0
✓	✗	bidirectional	93.0	59.2	82.6
✗	✓	bidirectional	93.1	59.7	82.3
✗	✗	bidirectional	93.0	58.9	82.3
✓	✓	up→down	93.2	59.1	82.6
✓	✗	up→down	93.1	58.8	82.3
✗	✓	up→down	93.0	58.9	82.4
✗	✗	up→down	92.7	58.4	82.2
✓	✓	down→up	93.1	58.9	82.5
✓	✗	down→up	93.1	58.0	82.3
✗	✓	down→up	92.9	59.3	82.1
✗	✗	down→up	92.7	58.6	82.1

4.4.4. Analysis of distillation hyper-parameter

In Eq. (15), a hyper-parameter is employed to regulate the relative influence of the distillation loss and distance loss. As illustrated in Fig. 4, this hyper-parameter significantly affects the accuracy of few-shot recognition. Notably, the recognition accuracy peaks near 1 for both SSV2-Small and Kinetics datasets. Therefore, in all other experiments, we maintain the hyper-parameter at a value of 1.

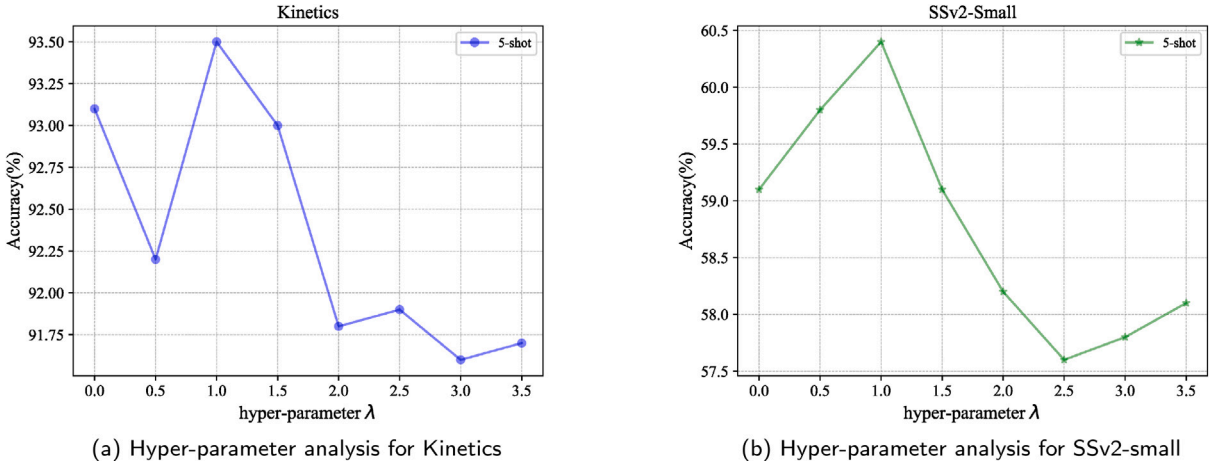
4.5. Distribution comparison

To illustrate how labels and temporal context contribute to maintaining class consistency across different contexts, we utilize t-SNE [53] to visualize the data distribution for Kinetics and SSv2 datasets under 5-way 5-shot settings.

In Figs. 5 and 6, we present the data distribution of CLIP-FSAR alongside the single view of our model, which incorporates visual, textual, and temporal context information, for Kinetics and SSv2-Small. Irrespective of whether it is the LTC view or the GTC view, these figures vividly depict that incorporating text and temporal context enhances inter-class discriminability while maintaining stable intra-class distribution. Moreover, our model's views exhibit fewer outliers compared to CLIP-FSAR.

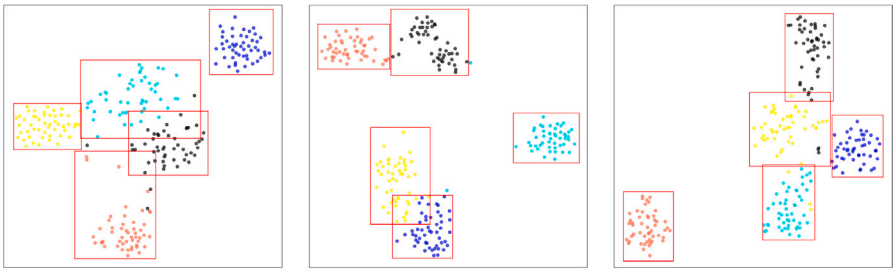
4.6. Accuracy comparison of different classes

In comparison to CLIP-FSAR [6] on the Kinetics dataset, we conducted meta-testing with 10 randomly selected classes from 24 classes under the 5-way 5-shot setting. The results, depicted in Figs. 7(a) and 7(b), exhibit significant accuracy enhancements for CLIP-MDMF across diverse classes. Notably, in Kinetics, the "shearing sheep" class showcases the most substantial improvement, with CLIP-MDMF achieving an accuracy of 98.6%, while CLIP-FSAR attains 92.3%. Similarly, on SSv2-Small, our model outperforms CLIP-FSAR in accuracy for most classes. For instance, the accuracy for the class "Poking a stack of something so the stack collapses" significantly rises from 61.0% with CLIP-FSAR to 81.0% with CLIP-MDMF. These findings underscore the efficacy of CLIP-MDMF in enhancing accuracy, particularly for specific action classes. Note: The class labels we select for SSV2-small are as follows: "Dropping something into something", "Letting something roll up a slanted surface, so it rolls back down", "Opening something", "Poking a stack of something so the stack collapses", "Pushing something off of something", "Putting something next to something", "Putting something on the edge of something so it is not supported and falls down", "Scooping something up with something", "something falling like a feather or paper', "Unfolding something".



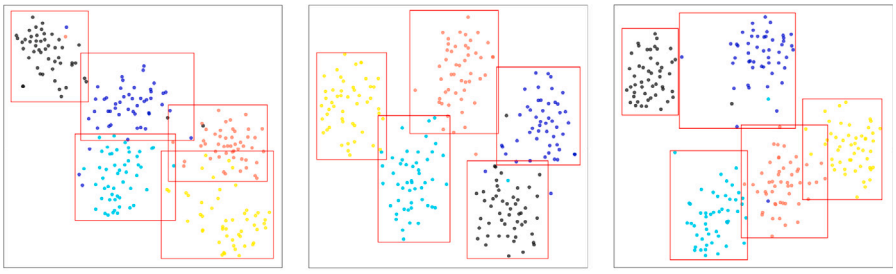
(a) Hyper-parameter analysis for Kinetics (b) Hyper-parameter analysis for SSv2-small

Fig. 4. Hyper-parameter λ for accuracy.



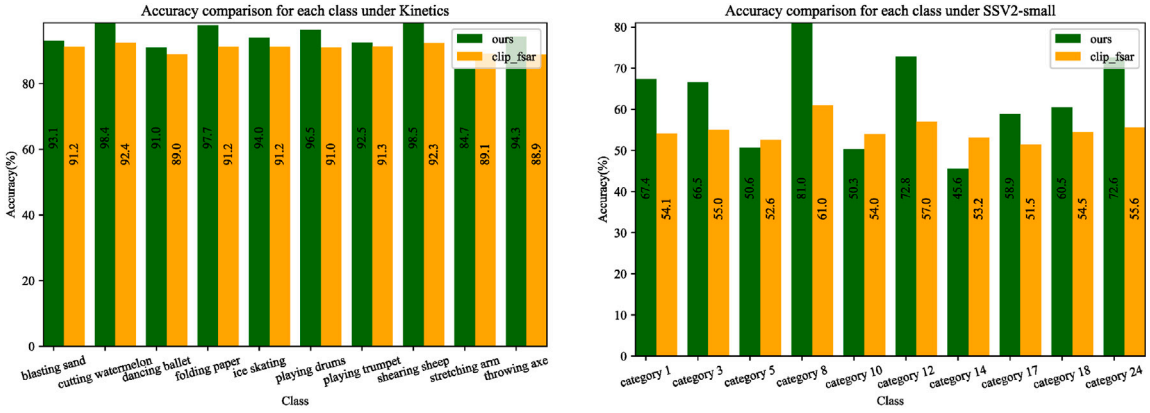
(a) CLIP-FSAR (b) Ours-LTC view (c) Ours-GTC view

Fig. 5. Distribution comparison on Kinetics.



(a) CLIP-FSAR (b) Ours-LTC view (c) Ours-GTC view

Fig. 6. Distribution comparison on SSv2.



(a) Accuracy under Kinetics (b) Accuracy under SSv2-small

Fig. 7. The illustration shows the comparison.

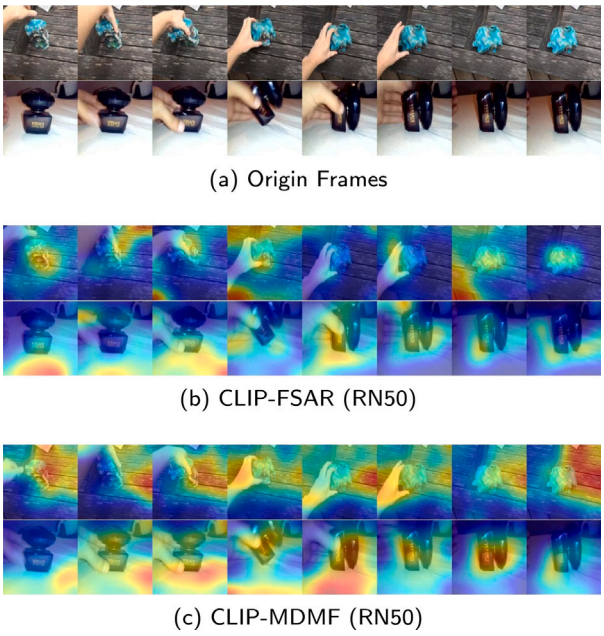


Fig. 8. Attention visualization for SSv2 based on CLIP-RN50 in 5-way 5-shot setting.

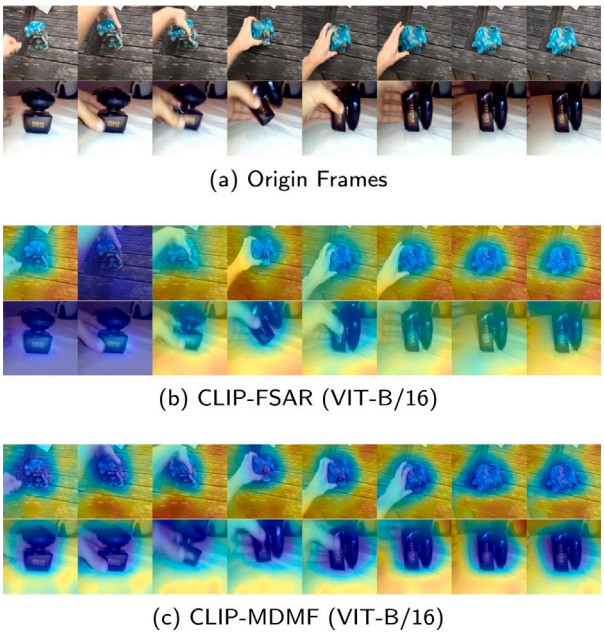


Fig. 10. Attention visualization for SSv2 based on VIT-B/16 in 5-way 5-shot setting.

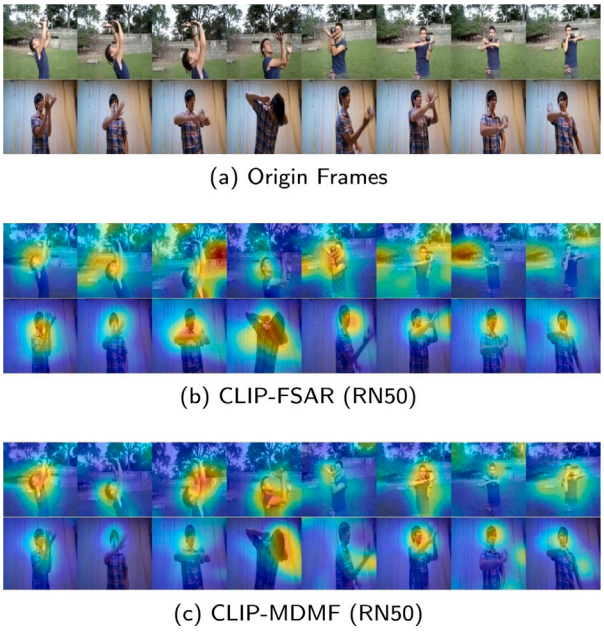


Fig. 9. Attention visualization for Kinetics based on CLIP-RN50 in 5-way 5-shot setting.

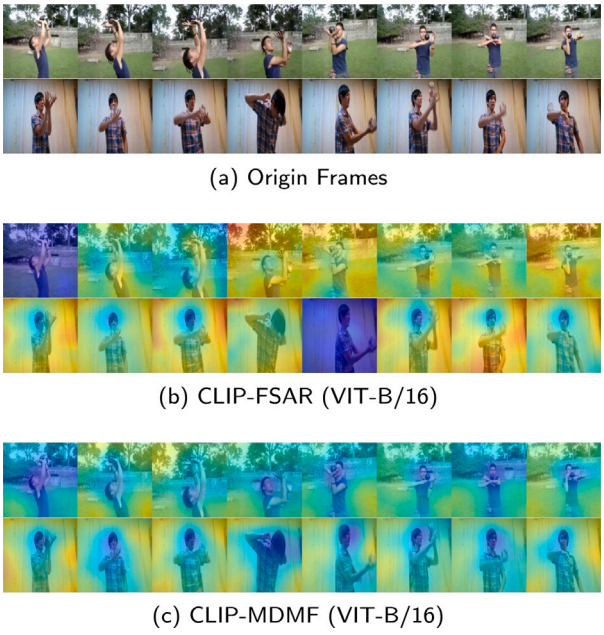


Fig. 11. Attention visualization for Kinetics based on VIT-B/16 in 5-way 5-shot setting.

4.7. Comparison for attention visualization

To further study the features, attention visualizations of our model are performed and compared with the attention visualizations [61] of CLIP-FSAR. We use the RN50 and VIT-B/16 as our backbone. From Figs. 8 to 11, in each figure, according to the RGB image sequence in sub-figure (a), the attention visualizations of CLIP-FSAR in sub-figure (b) are compared with the attention visualizations of our model in sub-figure (c). For the sequence pair in each sub-figure, the first one is the support, and the second one is the query.

Attention under the CLIP(RN50). In SSv2-small, for the action category "Laying something on the table on its side, not upright", visualizations in Fig. 8 show that compared with CLIP-FSAR, our CLIP-MDMF effectively directs attention towards action-related backgrounds while minimizing focus on unrelated elements. Moreover, the attention is notably more precise. In Fig. 9, showcasing CLIP-MDMF's attention on Kinetics with the action category "contact juggling", we observe similar accuracy enhancements. The attention distribution demonstrates the reduced emphasis on irrelevant backgrounds, distinguishing it from CLIP-FSAR.

Attention under the CLIP(VIT-B/16). In Fig. 10, depicting the action class "Laying something on the table on its side, not upright", our model's attention exhibits enhanced accuracy, accompanied by more coherent sequence attentions. In Fig. 11, showcasing the attention visualization of our CLIP-MDMF on Kinetics under the 5-way 5-shot setting with the action class "contact juggling", notable differences emerge. Our model's attention spans the entire moving entity, contrasting with FSAR's attention, which concentrates solely on a portion of it.

Additionally, the sequence attentions of our model demonstrate greater coherence.

5. Scalability discussion for further work

In our work, the modalities include RGB vision and text. In the future, we may explore further extensions to modalities, such as optical flow, heatmap, and audio modalities, to enhance the robustness and versatility of our model. The modality data concatenation proposed in this paper, as well as the Cross-Transformer for concatenated data from various modalities, is also applicable to these extended modalities. Additionally, in multi-view distillation, if we handle data from multiple modalities simultaneously, the conditions for reliable and unreliable views will become more complex. These areas are worth continuing to explore to advance the field and improve model performance. In this paper, the modalities include RGB vision and text. In the future, we may explore further extensions to modalities, such as optical flow, heatmap, and audio modalities, to enhance the robustness and versatility of our model. The modality data concatenation proposed in this paper, as well as the Cross-Transformer for concatenated data from various modalities, is also applicable to these extended modalities. Additionally, in multi-view distillation, if we handle data from multiple modalities simultaneously, the conditions for reliable and unreliable views will become more complex. These areas are worth continuing to explore to advance the field and improve model performance.

6. Conclusion

In this paper, we aim to fully leverage both textual and visual modalities in few-shot action recognition. We utilize CLIP as the backbone for processing both visual inputs and label text. To obtain the probability prompt embedding for the query, we introduce the Probability Prompting Selector (PPS), which employs matching scores and uniform sampling. Within the global context view or local context view, we integrate the prompt embedding with the visual embedding and temporal context using the Multi-modal Fusion Encoder (MMFE), allowing us to capture both local and global temporal contexts effectively. Furthermore, we incorporate distance fusion and mutual distillation techniques to facilitate mutual information exchange between views, thereby enhancing their reliability.

CRediT authorship contribution statement

Fei Guo: Writing – original draft, Software, Project administration, Conceptualization, Formal analysis, Methodology, Visualization. **YiKang Wang:** Validation, Data curation, Software. **Han Qi:** Validation, Investigation, Writing – review & editing. **Wenping Jin:** Investigation, Validation. **Li Zhu:** Supervision, Resources, Funding acquisition. **Jing Sun:** Resources, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work is supported by the National Key Research and Development Program (Grant No. 2019YFB2102500) and the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2021A15150 11913).

References

[1] L. Zhu, Y. Yang, Label independent memory for semi-supervised few-shot video classification, IEEE Trans. Pattern Anal. Mach. Intell. 44 (1) (2020) 273–285.
[2] L. Zhang, X. Chang, J. Liu, M. Luo, M. Prakash, A.G. Hauptmann, Few-shot activity recognition with cross-modal memory network, Pattern Recognit. 108 (2020) 107348.
[3] Y. Fu, L. Zhang, J. Wang, Y. Fu, Y.-G. Jiang, Depth guided adaptive meta-fusion network for few-shot video recognition, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1142–1151.
[4] X. Ni, H. Wen, Y. Liu, Y. Ji, Y. Yang, Multimodal prototype-enhanced network for few-shot action recognition, 2022, arXiv preprint arXiv:2212.04873.
[5] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[6] X. Wang, S. Zhang, J. Cen, C. Gao, Y. Zhang, D. Zhao, N. Sang, CLIP-guided prototype modulating for few-shot action recognition, 2023, arXiv preprint arXiv:2303.02982.
[7] Y. Wanyan, X. Yang, C. Chen, C. Xu, Active exploration of multimodal complementarity for few-shot action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 6492–6502.
[8] C. Simon, P. Koniusz, R. Nock, M. Harandi, Adaptive subspaces for few-shot learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4136–4145.
[9] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, Adv. Neural Inf. Process. Syst. 30 (2017).
[10] F. Sung, Y. Yang, L. Zhang, T. Xiang, P.H. Torr, T.M. Hospedales, Learning to compare: Relation network for few-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1199–1208.
[11] P. Singh, P. Mazumder, Dual class representation learning for few-shot image classification, Knowl.-Based Syst. 238 (2022) 107840.
[12] J. Zhou, Q. Lv, C.Y.-C. Chen, Dynamic concept-aware network for few-shot learning, Knowl.-Based Syst. 258 (2022) 110045.
[13] S. Deng, Z. Guo, D. Teng, B. Lin, D. Chen, T. Jia, H. Wang, Self-relation attention networks for weakly supervised few-shot activity recognition, Knowl.-Based Syst. (2023) 110720.
[14] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 1126–1135.
[15] Y. Shao, W. Wu, X. You, C. Gao, N. Sang, Improving the generalization of MAML in few-shot classification via bi-level constraint, IEEE Trans. Circuits Syst. Video Technol. (2022) 1, http://dx.doi.org/10.1109/TCSVT.2022.3232717.
[16] S. Ravi, H. Larochelle, Optimization as a model for few-shot learning, in: International Conference on Learning Representations, 2017.
[17] Y. Feng, J. Chen, J. Xie, T. Zhang, H. Lv, T. Pan, Meta-learning as a promising approach for few-shot cross-domain fault diagnosis: Algorithms, applications, and prospects, Knowl.-Based Syst. 235 (2022) 107646.
[18] A.A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, R. Hadsell, Meta-learning with latent embedding optimization, 2018, arXiv preprint arXiv:1807.05960.
[19] Y. Zheng, X. Zhang, Z. Tian, W. Zeng, S. Du, Detach and unite: A simple meta-transfer for few-shot learning, Knowl.-Based Syst. 277 (2023) 110798.
[20] Z. Chen, Y. Fu, Y. Zhang, Y. Jiang, X. Xue, L. Sigal, Semantic feature augmentation in few-shot learning. arxiv 2018, 2018, arXiv preprint arXiv:1804.05298.
[21] A.J. Ratner, H. Ehrenberg, Z. Hussain, J. Dunnmon, C. Ré, Learning to compose domain-specific transformations for data augmentation, Adv. Neural Inf. Process. Syst. 30 (2017).
[22] L. Perez, J. Wang, The effectiveness of data augmentation in image classification using deep learning, 2017, arXiv preprint arXiv:1712.04621.
[23] F. Pahde, M. Puscas, T. Klein, M. Nabi, Multimodal prototypical networks for few-shot learning, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 2644–2653.
[24] Z. Dang, M. Luo, C. Jia, C. Yan, X. Chang, Q. Zheng, Counterfactual generation framework for few-shot learning, IEEE Trans. Circuits Syst. Video Technol. (2023) 1, http://dx.doi.org/10.1109/TCSVT.2023.3241651.
[25] S. Shao, Y. Wang, B. Liu, W. Liu, Y. Wang, B. Liu, FADS: Fourier-augmentation based data-shunting for few-shot classification, IEEE Trans. Circuits Syst. Video Technol. (2023) 1, http://dx.doi.org/10.1109/TCSVT.2023.3292519.
[26] L. Zhu, Y. Yang, Compound memory networks for few-shot video classification, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 751–766.
[27] K. Cao, J. Ji, Z. Cao, C.-Y. Chang, J.C. Niebles, Few-shot video classification via temporal alignment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10618–10627.
[28] M. Müller, Dynamic time warping, Inf. Retr. Music Motion (2007) 69–84.
[29] T. Perrett, A. Masullo, T. Burghardt, M. Mirmehdi, D. Damen, Temporal-relational crosstransformers for few-shot action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 475–484.

[30] A. Thatipelli, S. Narayan, S. Khan, R.M. Anwer, F.S. Khan, B. Ghanem, Spatio-temporal relation modeling for few-shot action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19958–19967.
[31] F. Guo, L. Zhu, Y. Wang, J. Sun, Task-specific alignment and multiple-level transformer for few-shot action recognition, Neurocomputing (2024) 128044.
[32] J. Wu, T. Zhang, Z. Zhang, F. Wu, Y. Zhang, Motion-modulated temporal fragment alignment network for few-shot action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 9151–9160.
[33] X. Wang, S. Zhang, Z. Qing, C. Gao, Y. Zhang, D. Zhao, N. Sang, MoLo: Motion-augmented long-short contrastive learning for few-shot action recognition, 2023, arXiv preprint arXiv:2304.00946.
[34] S. Kumar Dwivedi, V. Gupta, R. Mitra, S. Ahmed, A. Jain, Protogan: Towards few shot learning for action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.
[35] Y. Xian, B. Korbar, M. Douze, L. Torresani, B. Schiele, Z. Akata, Generalized few-shot video classification with video retrieval and feature generation, IEEE Trans. Pattern Anal. Mach. Intell. 44 (12) (2021) 8949–8961.
[36] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).
[37] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[40] W. Park, D. Kim, Y. Lu, M. Cho, Relational knowledge distillation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3967–3976.
[41] F. Tung, G. Mori, Similarity-preserving knowledge distillation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1365–1374.
[42] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, 2015, arXiv preprint arXiv:1503.02531.
[43] C. Shen, X. Wang, J. Song, L. Sun, M. Song, Amalgamating knowledge towards comprehensive classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 3068–3075.
[44] Y. Zhang, T. Xiang, T.M. Hospedales, H. Lu, Deep mutual learning, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2018.
[45] S. Albanie, A. Nagrani, A. Vedaldi, A. Zisserman, Emotion recognition in speech using cross-modal transfer in the wild, in: Proceedings of the 26th ACM International Conference on Multimedia, 2018, pp. 292–301.
[46] C. Lea, R. Vidal, A. Reiter, G.D. Hager, Temporal convolutional networks: A unified approach to action segmentation, in: Computer Vision–ECCV 2016 Workshops: Amsterdam, the Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14, Springer, 2016, pp. 47–54.
[47] K. Soomro, A.R. Zamir, M. Shah, UCF101: A dataset of 101 human actions classes from videos in the wild, 2012, arXiv preprint arXiv:1212.0402.
[48] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, HMDB: a large video database for human motion recognition, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 2556–2563.
[49] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.
[50] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. 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, 2017, pp. 5842–5850.
[51] H. Zhang, L. Zhang, X. Qi, H. Li, P.H. Torr, P. 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, Springer, 2020, pp. 525–542.
[52] D. Kingma, J. Ba, Adam: A method for stochastic optimization, Comput. Sci. (2014).
[53] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. Van Gool, Temporal segment networks: Towards good practices for deep action recognition, in: European Conference on Computer Vision, Springer, 2016, pp. 20–36.
[54] X. Wang, S. Zhang, Z. Qing, M. Tang, Z. Zuo, C. Gao, R. Jin, N. Sang, Hybrid relation guided set matching for few-shot action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19948–19957.
[55] S. Li, H. Liu, R. Qian, Y. Li, J. See, M. Fei, X. Yu, W. Lin, TA2N: Two-stage action alignment network for few-shot action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 1404–1411.
[56] S. Zheng, S. Chen, Q. Jin, Few-shot action recognition with hierarchical matching and contrastive learning, in: European Conference on Computer Vision, Springer, 2022, pp. 297–313.
[57] S. Liu, M. Jiang, J. Kong, Multidimensional prototype refactor enhanced network for few-shot action recognition, IEEE Trans. Circuits Syst. Video Technol. 32 (10) (2022) 6955–6966, http://dx.doi.org/10.1109/TCSVT.2022.3175923.
[58] X. Wang, W. Ye, Z. Qi, G. Wang, J. Wu, Y. Shan, X. Qie, H. Wang, Task-aware dual-representation network for few-shot action recognition, IEEE Trans. Circuits Syst. Video Technol. (2023) 1, http://dx.doi.org/10.1109/TCSVT.2023.3262670.
[59] X. Wang, W. Ye, Z. Qi, X. Zhao, G. Wang, Y. Shan, H. Wang, Semantic-guided relation propagation network for few-shot action recognition, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 816–825.
[60] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al., Matching networks for one shot learning, Adv. Neural Inf. Process. Syst. 29 (2016).
[61] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.