Efficient Few-Shot Action Recognition via Multi-Level Post-Reasoning

Cong Wu[1], Xiao-Jun Wu[1⋆], Linze Li[1], Tianyang Xu[1], Zhenhua Feng[1,2,3], and Josef Kittler[2,3]

[1] School of Artificial Intelligence and Computer Science, Jiangnan University, China
[2] Centre for Vision, Speech and Signal Processing, University of Surrey, UK
[3] School of Computer Science and Electronic Engineering, University of Surrey, UK
{congwu,linze.li}@stu.jiangnan.edu.cn;
{wu_xiaojun,tianyang.xu,fengzhenhua}@jiangnan.edu.cn;
j.kittler@surrey.ac.uk

Abstract. The integration with CLIP (Contrastive Vision-Language Pre-training) has significantly refreshed the accuracy leaderboard of FSAR (Few-Shot Action Recognition). However, the trainable overhead of ensuring that the domain alignment of CLIP and FSAR is often unbearable. To mitigate this issue, we present an **E**fficient **M**ulti-Level **P**ost-Reasoning **Net**work, namely **EMP-Net**. By design, a post-reasoning mechanism is proposed for domain adaptation, which avoids most gradient backpropagation, improving the efficiency; meanwhile, a multi-level representation is utilised during the reasoning and matching processes to improve the discriminability, ensuring effectiveness. Specifically, the proposed EMP-Net starts with a skip-fusion involving cached multi-stage features extracted by CLIP. After that, the fused feature is decoupled into multi-level representations, including global-level, patch-level, and frame-level. The ensuing spatiotemporal reasoning module operates on multi-level representations to generate discriminative features. As for matching, the contrasts between text-visual and support-query are integrated to provide comprehensive guidance. The experimental results demonstrate that EMP-Net can unlock the potential performance of CLIP in a more efficient manner. The code and supplementary material can be found at https://github.com/cong-wu/EMP-Net.

Keywords: Image to Video Adaption · Contrastive Vision-Language Pre-training · Efficient Few-Shot Action Recognition

1 Introduction

Collecting sufficient training samples is often impractical in action recognition, as collecting and annotating videos is extremely challenging and demands considerable time and labour consuming. Motivated by this difficulty, FSAR has attracted widespread attention in the community. The dominant paradigm in

⋆ Corresponding Author.

Table 1: A comparison with CLIP-FSAR under different training paradigms on 2 GeForce RTX 3090 GPU with batchsize as 2 on SSv2-Small (5-way 1-shot). OOM means CUDA out of memory error.

Method	Backbone	GPU Memory	Accuracy
CLIP-FSAR [36]	Trainable CLIP	>24,268MiB (OOM)	-
	Adapter CLIP	>24,268MiB (OOM)	-
	Frozen CLIP	6,799MiB	50.1%
EMP-Net	Frozen CLIP	8,565MiB	57.1%

FSAR is the metric learning framework, which uses the distance between corresponding samples measured in a high-dimensional space as the basis for classification [1, 2]. Unlike conventional training paradigms, approaches based on metric learning introduce another auxiliary dataset with a non-overlapped label space as compared with the existing target dataset. Subsequently, the knowledge learned from the auxiliary dataset is transferred to the target dataset. Existing methods use various means to improve the feature discriminability [53] or the comprehensiveness of measurement [13, 47], thereby enhancing the reliability of matching.

Recently, Wang *et al*. introduced CLIP-FSAR [36]. This pioneering work integrates CLIP [27] with FSAR, establishing a State-of-the-Art (SOTA) solution. In CLIP-FSAR, pre-trained CLIP encoders act as basic text and visual feature extractors. Combining with temporal reasoning blocks, it infers the final labels using the metric learning paradigm. Despite significant performance improvement compared with other methods, the fine-tuning of the backbone network incurs substantial training costs. As shown in Table 1, the training of CLIP-FSAR necessitates high-memory hardware, posing challenges in accessibility for most scenarios. The first effort aimed at improving training efficiency is to transform the backbone from trainable CLIP into adapter CLIP [24, 45] by allowing only the newly introduced parameters to be updated during the training stage. However, as the newly introduced parameters spread throughout the entire backbone, gradient backpropagation still occurs, resulting in limited savings in the training overhead. Another strategy is to freeze the backbone completely [39]. However, as shown in Table 1, the efficiency gain is invariably at the expense of reduced accuracy, owing to the failure to bridge the gap between the source domain and the target task. Recent study [50] shows that powerful representation models can be constructed by mining the features of frozen feature extractors. In summary, although CLIP has the potential to play a very useful role in FSAR, the existing methods struggle to achieve satisfactory performance in terms of both effectiveness and efficiency.

To overcome the above challenges, we present EMP-Net by leveraging the latest developments in vision-language pre-training for the benefit of the FSAR task in an efficient and effective manner. The main innovation of our EMP-Net is two-fold: post-reasoning and multi-level modelling. For post-reasoning, EMP-Net ensures that all trainable structures are located after the feature extractor, which means that the reasoning operations are based on the cached extracted

features. Also, different levels of features often exhibit complementarity, which has been proved repeatedly [17,53]. To leverage this, EMP-Net constructs multi-level representations based on the global, patch, and frame perspectives, thereby improving the discriminability and reliability of subsequent reasoning and matching tasks. The above two innovations work together to greatly improve training efficiency while ensuring accuracy.

Following the metric learning framework, the proposed EMP-Net consists of two main components: image2video adaptation and matching metric. Notably, using the cached visual features extracted from different layers of CLIP, we construct a skip-fusion module that performs spatiotemporal reasoning, which maximises the utilisation of information at different granularities. The obtained features are further deconstructed to obtain representations at the global, patch, and frame levels. By combining interaction spatiotemporal reasoning and text-guided temporal reasoning based on multi-level representations, we obtain rich and comprehensive visual clues. As for the matching process, we introduce a multi-level matching algorithm that learns the associations between action labels and given videos, as well as the associations between the support and query videos. We evaluate EMP-Net on multiple datasets, including a temporal-related dataset Something-Something V2 (Small&Full), and static-related datasets such as HMDB-51, UCF-101, as well as Kinetics-400. The experimental results demonstrate that our method surpasses or achieves SOTA performance with extremely low training costs.

The key innovations of EMP-Net can be summarised as follows:

- We propose a novel model for efficient few-shot action recognition.
- The image2video adaptation mechanism is designed for knowledge transfer from CLIP to FSAR, which includes a skip-fusion for initial alignment, and the following interaction reasoning and text-guided reasoning for spatiotemporal modelling.
- Multi-level features are applied in the reasoning and matching processes to obtain more discriminative representations.
- To the best of our knowledge, this is the first solution that achieves a good balance between training costs and accuracy.

2 Related Work

2.1 Few-Shot Learning

Few-shot learning aims to derive effective representations from extremely limited annotated data [40]. Existing few-shot learning methods follow a customised transfer learning paradigm, where knowledge is acquired from a category-agnostic auxiliary training set and then transferred to the current test set. Different from conventional action recognition tasks, the label spaces of training and test sets are non-shared. Also, the accessible labelled samples are extremely limited during the training process. According to [20], those methods can be broadly categorised into three types: (1) Non-episodic based methods [5, 7]: Non-episodic few-shot

learning involves two stages. First, the entire network is trained on the auxiliary data, and then only the classifier is fine-tuned on the target set. (2) Meta-learning based methods [9, 49]: Meta-learning entails the training of a base learner and a meta learner. The former is trained on the auxiliary dataset, while the latter is trained on the target set. (3) Metric learning based methods [19, 30, 32]: In this framework, the current set is first divided into support and query sets. Then, the network is trained to classify the query sample into the corresponding category based on the distance between the query and support set samples.

2.2 Action Recognition

Action recognition identifies the category of action depicted in a video clip [14]. Unlike image recognition, action recognition requires the understanding of a continuous sequence of frames, which necessitates the modelling of temporal information. In the early stages, most feature descriptors, such as 3D-SIFT [28], extended SURF [41], used for action recognition are extensions of classic 2D descriptors. Additionally, some motion descriptors, such as MBH [6], have been proposed. More recently, deep learning frameworks also followed this paradigm, either exploring spatiotemporal feature modelling, or constructing motion descriptors, or a combination of both. For example, Simonyan *et al*. [29] proposed a two stream network, which includes a spatial branch for spatial modelling and a temporal branch to capture the motion information based on optical flow. In TSM (Temporal Shift Module) [21], the temporal shift operation is introduced to endow 2D CNNs with the temporal awareness ability. This operation has also been applied to a Transformer-based framework by Xiang *et al*. [44]. Recently, topics surrounding unsupervised learning have also received widespread attention [26, 42].

2.3 Few-Shot Action Recognition

Few-shot learning has garnered considerable attention and inspired research across various tasks [15, 32], including action recognition [51]. Most FSAR approaches follow the metric learning paradigm. In TARN [2], an attention-based relation module was proposed to measure the distance among aligned segments. Recently, several methods have been proposed to explore the matching metrics. For example, OTAM [3] utilised the knowledge of event ordering in a sequence to create a frame-based matching strategy; HyRSM [38] built a task-specific matching algorithm to find the most reasonable pairs among the corresponding support and query samples; Wang *et al*. [37] designed a novel framework, MOLO, that combines frame-frame matching and frame-video matching, furthermore, an additional motion auto-decoder was proposed to enhance the discrimination.

Following the above methods, CLIP-FSAR [36] is the first attempt that leverages the powerful modelling capability of CLIP to represent videos and their corresponding labels, resulting in a significant performance boost. However, the huge computational cost associated with fine-tuning CLIP cannot be overlooked. In addressing this dilemma, approaches like D^2ST-Adapter [24] and

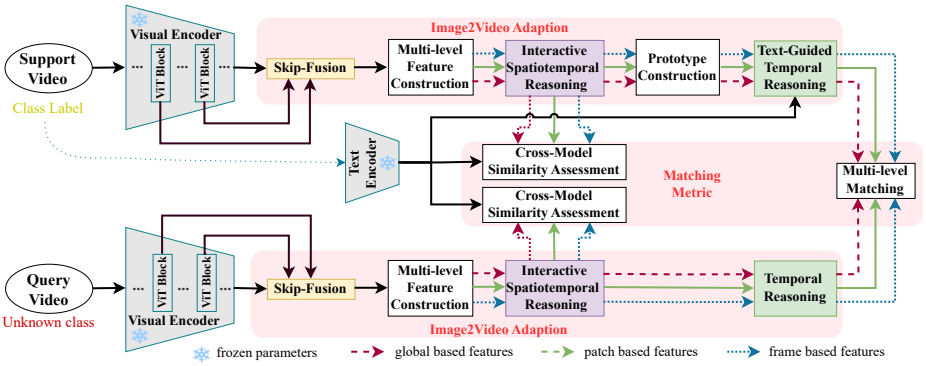

Fig. 1: The proposed EMP-Net has three main components: frozen CLIP for visual and textual features extraction, image-to-video adaption for spatiotemporal reasoning, and the matching metric for classification.

MA-CLIP [45] attempted to borrow the concept of adapter [23]. Undoubtedly, this process greatly reduces the number of trainable parameters. However, since there are newly introduced parameters that need to be updated at the front of a network, the backpropagation of gradients still exists, which will cause most of the training overhead to still be unavoidable. There are also methods [39] that use the completely frozen encoder to extract basic information and, on this basis, employ additional trainable modelling modules for further reasoning. However, this approach, while highly efficient in training, comes with a significant compromise in accuracy. In contrast to these methods, our approach achieves both efficiency and effectiveness, surpassing all other paradigms significantly.

3 The Proposed Method

3.1 The Overall Paradigm

Problem Definition. FSAR aims at recognising the classes based on very limited labelled samples. Following the typical metric learning paradigm, we construct a support set and a query set during each training iteration. Then the network is trained to classify the query samples into one of the classes from the support set, taking the support samples as a reference. In the N-way K-shot task, N denotes the number of selected action labels, and K is the number of samples corresponding to each category in the support set. The query set contains several unknown samples belonging to given categories. The optimisation is based on the premise that the query sample is in the same category as the support sample closest to it. Then the network is trained to enable query samples to be classified into the correct categories, based on the measured distance to given support samples. Accordingly, the matching ability learned on the training set is transferred to the test set for evaluation.

Overall Architecture. For each iteration, as shown in Fig. 1, our model starts with the frozen visual encoder to extract the basic visual information from the given videos, as well as the frozen text encoder to extract the textual information from the given labels. The extracted features are then used by the image2video adaption module for spatiotemporal reasoning. For any input video, the multi-level visual representations are generated by the image2video adaption module. The modelling process for support and query videos is basically the same. The first difference is that the text label corresponding to the query is unknown, so there is no interaction between the corresponding textual information and query information. Also, the prototype construction is only used on the support branch to merge the features from the same class, which will be explained later in Sec. 3.2. Based on the extracted textual and visual representations, the matching metric is defined. Specifically, the matching process involves comparing visual information and textual information, as well as comparing the visual information conveyed by the support and query samples respectively.

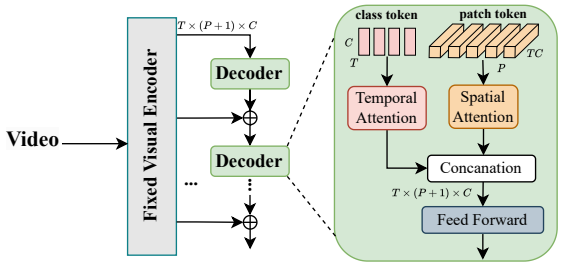

Fig. 2: Skip-Fusion combines the features of different stages by progressive fusion which is implemented by an efficient spatial-temporal reasoning decoder.

3.2 Image2Video Adaption

Skip-Fusion. Existing methods [36,37] usually perform reasoning based on the feature obtained from the last layer of the visual encoder. Here we argue that this paradigm is not necessarily suitable for FSAR because it is difficult to provide sufficient discriminability. This problem is exacerbated when using frozen CLIP as the backbone. Motivated by this, we devise a novel block, named skip-fusion, leveraging the cached features from various stages of the frozen visual encoder to generate a more dependable representation. On the one hand, features at different levels focus on different granularities, and therefore a combination of these features should help to improve the discriminability of the network. On the other hand, the adaptation conducted on features across different stages simultaneously facilitates the transformation from CLIP to the current task domain, thereby endowing the currently frozen encoder with heightened flexibility. This idea can be considered as side-tuning, which is an efficient transfer learning

paradigm that has been proven effective in several tasks [10, 33, 48]. Different from previous works, a more efficient and compact module will be constructed here.

As illustrated in Fig. 2, we design a progressive reasoning and fusion module that takes features from shallow to deep layers into consideration. The details of feature selection will be discussed in Sec. 4.2. For any selected feature $fv_l \in \mathbb{R}^{T\times(P+1)\times C}$, P represents the number of patches, T is the number of frames, and C is the channel dimension, l is the index of the selected candidate. We fist fuse fv_l with the previous decoded feature fv_{l-1}^d to generate $fv_l^{'}$, and then send the obtained feature into the decoder for parallel spatiotemporal reasoning. If fv_{l-1}^d does not exist, $fv_l^{'} = fv_l$. As for the operation inside the decoder, first, we decompose the class and patch tokens from the current feature. As they were originally designed [8], class tokens contain global information, and patch tokens provide the local spatial clues. Here we input the former into the temporal transformer layer and feed the latter into the spatial transformer layer. The result is further fused by concatenation along the channel dimension. The fused feature is then passed through a feed-forward layer to obtain the output. The operation inside Decoder can be summarised as,

$$\begin{aligned} fv_l^d &= \text{Decoder}(fv_l^{'}), \\ &= \text{FFD}(\text{Concat}[\text{T-MHA}(fv_l^{'}[:,0]), \text{S-MHA}(fv_l^{'}[:,1:])]). \end{aligned} \tag{1}$$

T-MHA and S-MHA are the temporal and spatial multi-head attention operations, Concat is concatenation along the channel dimension, and FFD represents the feed-forward operation. The LayerNorm and Residual connection operations follow the design of ViT [8], which are omitted here for brevity, similar with Eq. (2) and Eq. (3). Compared with performing spatial and temporal operations serially, this design is undoubtedly more intuitive and efficient.

Interactive Spatiotemporal Reasoning. As a prerequisite to reasoning, we construct multi-level features as shown in Fig. 1. The key innovation is decomposing the final feature from skip-fusion module into global, patch, and frame representations. We take the class tokens and patch tokens to directly build global representation $fv^g \in \mathbb{R}^{T\times C}$ and patch representation $fv^p \in \mathbb{R}^{T\times P\times C}$ respectively. As for the frame-based representation $fv^f \in \mathbb{R}^{T\times C}$, we construct it by employing an average pooling operation on the spatial dimension of patch representation. This step provides a more comprehensive perspective on the different representations for the subsequent reasoning and matching by exploiting their complementarity. Subsequent reasoning is an extremely critical process, which is something that previous work failed to accomplish [53].

As shown in the left of Fig. 3, the customised reasoning operation is performed on the extracted multi-level features. Specifically, for the global feature, we simply perform temporal reasoning along the temporal dimension. Subsequently, the resulting feature is concatenated with the patch-based feature along the spatial dimension, and the newly generated feature is sent into a spatial reasoning block. We fuse the frame-based feature with the pooled patch feature and then use it as the feature for reasoning along the temporal dimension. This

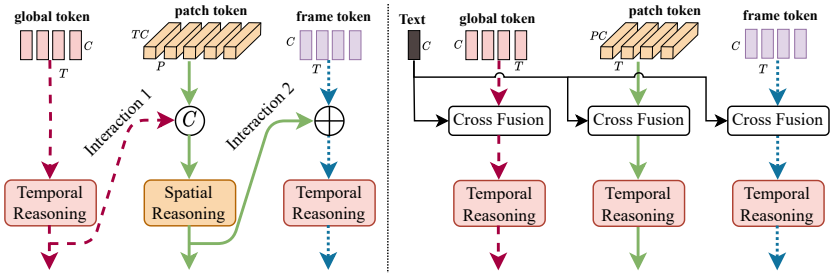

Fig. 3: The Image2Video adaption mechanism. Interactive Spatiotemporal Reasoning (left) and Text-Guided Temporal Reasoning (right). The above reasoning modules are conducted on multi-level representations, therefore greatly increasing the discriminability.

process can be summarised as follows,

$$\begin{aligned} iv^g &= \text{FFD}(\text{T-MHA}(fv^g)), \\ iv^p &= \text{FFD}(\text{S-MHA}(\text{Concat}[iv^g, fv^p]))[:, 1:], \\ iv^f &= \text{FFD}(\text{T-MHA}(\text{S-GAP}(iv^p) + fv^f)). \end{aligned} \tag{2}$$

S-GAP represents the global average pooling on the spatial dimension. The other symbols are similar with Eq. (1). The interactive mechanism invoked within the reasoning operation ensures that the complementary sources of information flows are fully aggregated.

Text-Guided Temporal Reasoning. Following the mainstream methods [30, 36], instead of retaining sample-based features for the support branch, we construct the category-based prototype representation by merging the video features from the same class. Based on given feature iv, the prototype representation $pv_n = \frac{1}{K}\sum_{k=1}^{K} iv_{kn}$, with n and k the index of categories and samples. To exploit the powerful visual-text modelling ability of CLIP fully, we investigate the effect of integrating text information into the current multi-level features corresponding to the support video. Specifically, the text information is first fused with multi-level features respectively, and then the output is sent into the temporal reasoning block, as shown in the right of Fig. 3. The aforementioned operation for each feature stream can be generally summarised as follows,

$$\begin{aligned} Q = \text{Sum}[pv, ft], K = V = \text{Concat}[pv, ft], \\ gv = \text{FFD}(\text{CT-MHA}(Q, K, V)), \end{aligned} \tag{3}$$

where $ft \in \mathbb{R}^C$ is the extracted textual feature from frozen text encoder, and CT-MHA is the cross temporal multi-head attention. It should be noted that since the dimensions of pv and ft are not consistent, we need to first expand or repeat the specific dimensions of ft before fusion. For example, for class-based and frame-based features, before the summation and concatenation, we expand the shape of ft in the temporal dimension. The resulting feature shape will be $T \times C$. A similar operation is applied to the token-based representation.

As for the feature in the query branch, we directly apply the temporal reasoning operation on it, where the operation consists of a temporal multi-head attention operation followed by a feed-forward operation.

3.3 Multi-level Matching Metric

Cross-Modal Similarity Assessment. As shown in Fig. 1, we measure the correlation between the text features and visual signals with the cosine similarity. Here we general denote the features after interactive reasoning as $\mathcal{IV}$, the textual features extracted by frozen text encoder as $\mathcal{FT}$. Then the correlation $\mathcal{D}_{\mathcal{VT}}$ can be obtained by Similarity$(\mathcal{IV}, \mathcal{FT})$. Given tensor A and B, the similarity matrix between A and B can be generated by $\frac{A \cdot B}{||A|| \times ||B||}$. Based on this definition, we can have $\mathcal{D}^g_{\mathcal{VT}}$, $\mathcal{D}^p_{\mathcal{VT}}$ and $\mathcal{D}^f_{\mathcal{VT}}$ for global, patch, and frame representations. Considering the characteristics of the different representations, we propose a novel selective mechanism to generate the probability distribution,

$$\begin{aligned}
\mathcal{P}^g_{\mathcal{VT}} &= \text{SoftMax}(\text{T-GAP}(\mathcal{D}^g_{\mathcal{VT}})), \\
\mathcal{P}^p_{\mathcal{VT}} &= \text{SoftMax}(\text{T-GAP}(\text{S-GAP}(\text{Top-K}(\mathcal{D}^p_{\mathcal{VT}})))), \\
\mathcal{P}^f_{\mathcal{VT}} &= \text{SoftMax}(\text{T-GAP}(\text{Top-M}(\mathcal{D}^f_{\mathcal{VT}}))).
\end{aligned} \tag{4}$$

Similar with Eq. (2), T-GAP is the global average pooling on the temporal dimension. Specifically, we directly use the pooling operation in the temporal dimension for the global-based similarity. For path-based similarity, we first select the TOP-K ($K = 49$) tokens and employ a spatial and temporal pooling operation. As for the frame-based similarities, the TOP-M ($M = 2$) selection is performed in the temporal dimension, followed by the pooling operation.

Support-Query Matching. Here we calculate the distance between support and query samples as

$$\mathcal{P}_{\mathcal{SQ}} = \text{SoftMax}(\text{OTAM}(\mathcal{GV}_{\mathcal{S}}, \mathcal{GV}_{\mathcal{Q}})). \tag{5}$$

OTAM [3] is adopted in our method as the temporal distance measurement. $\mathcal{GV}_{\mathcal{S}}$ and $\mathcal{GV}_{\mathcal{Q}}$ represent the final visual feature for the support and query videos.

Loss and Prediction. Considering the training set does not share the same label space as the test set, it would be inappropriate to classify the query videos into the original category labels. For this reason, we construct the local label set for loss calculation. Specifically, for the N categories involved in each iteration, we recode them from 0 to $N - 1$. The aim is to learn a more robust matching mechanism. The cross-entropy loss is used for the loss calculation.

$$\begin{aligned}
\mathcal{L} &= \mathcal{L}^g_{\mathcal{VT}} + \mathcal{L}^p_{\mathcal{VT}} + \mathcal{L}^f_{\mathcal{VT}} + \mathcal{L}^g_{\mathcal{SQ}} + \mathcal{L}^p_{\mathcal{SQ}} + \mathcal{L}^f_{\mathcal{SQ}}, \\
\mathcal{P} &= \mathcal{P}^g_{\mathcal{VT}} + \mathcal{P}^p_{\mathcal{VT}} + \mathcal{P}^f_{\mathcal{VT}} + \mathcal{P}^g_{\mathcal{SQ}} + \mathcal{P}^p_{\mathcal{SQ}} + \mathcal{P}^f_{\mathcal{SQ}}.
\end{aligned} \tag{6}$$

$\mathcal{L}_{\mathcal{VT}}$ represents the loss for cross-modal similarity assessment; and $\mathcal{L}_{\mathcal{SQ}}$ corresponds to the matching for support-query sets. The training is guided by the

Table 2: The impact of each component under 5-way 1-shot evaluation on SSv2-Small dataset. [1] Only a trainable fully connected layer is added after the Frozen-CLIP as the classification layer.

(a) Multi-Level& Post-Reasoning.

Multi-Level	Post-reasoning	SSv2-Small
✗	✗[1]	41.0
✗	✓	51.6
✓	✓	57.1

(b) Matching Metrics.

Text-Visual	Support-Query	SSv2-Small
✗	✓	56.1
✓	✓	57.1

(c) The details of Post-Reasoning.

Skip-Fusion	Interactive Reasoning	Text-guided Reasoning	SSv2-Small
✗	✗	✓	52.4
✗	✓	✓	54.1
✓	✗	✓	55.8
✓	✓	✓	57.1

sum of those loss functions on multi-level representations. Similarly, the final decision takes into account the similarity between vision and text, as well as the matching between the support and query sets. That is, the final classification score $\mathcal{P}$ can be obtained by Eq. (6). In the evaluation phase, only the score corresponding to the query set is counted during the calculation of $\mathcal{P}$.

4 Experiments

4.1 Datasets and Implementation Details

We select temporal-related datasets: SSV2 (Something-Something V2) Small&Full [11], and spatial-related datasets: K400 (Kinetics-400) [4], HMDB-51 [16], UCF-101 [31] for evaluation. The implementation details follow the CLIP-FSAR [36]. Please find more details in the supplementary material.

4.2 Ablation Studies

In this subsection, we first quantitatively analyse the merits of each innovation and the design details. We also conduct extensive comparisons with mainstream paradigms in terms of efficiency and effectiveness.

The impact of each component on accuracy. From Table 2a, with frozen CLIP as the feature extractor, and adding only a trainable fully connected layer as the classification layer, the accuracy is only 41% on SSv2-Small. When post-reasoning is introduced, the accuracy increases by 10.6%. After extending the whole framework to be multi-level, we achieve 57.1% in accuracy. These results demonstrate the effectiveness of post-reasoning and multi-level mechanisms. In Table 2b, we explore the impact of different matching metrics on the final performance. The results prove that the method of combining text-vision and support-query matching achieves the optimal result. We further evaluate the effectiveness

Table 3: Details of the designed module on SSV2-small, 5way-1shot.

(a) Multi-Level Representation.

Global Only	51.6
Global&Patch	55.4
Global&Patch&Frame	57.1

(b) Skip-Fusion.

Attention	Spatial Only	54.4
	Spatial&Temporal	57.1
Layer	{6,12}	55.4
	{4,8,12}	57.1
	{3,6,9,12}	56.2

(c) Interactive Spatiotemporal Reasoning.

Interactive-1	Interactive-2	
✗	✗	55.4
✗	✓	56.1
✓	✗	56.3
✓	✓	57.1

(d) Text-guided Temporal Reasoning.

Cross	No Q-Fusion	56.5
	No KV-Fusion	56.3
	Default Fusion	57.1
Without Text-Guided		55.5

(e) Matching Metrics.

	Selective Mechanism	
Text-Visual	✗	56.4
	✓	57.1
Support-Query	✗	57.1
	✓	56.5

of each component of post reasoning, as shown in Table 2c. We observe that the accuracy drops a lot with the skip-fusion and interactive reasoning (from 57.1% to 52.4%) removed. After adding the interactive reasoning, the performance improves by 1.7%; with skip-fusion, the performance improves another 1.7%. The final design incorporating these three modules achieves optimal performance.

The Design Details for Each Module. In this part, we analyse the design details, as summarised in Table 3. First of all, regarding the muti-level representation, from Table 3a, when we use the global representation only, same with previous mainstream methods, the performance is only 51.6%. When it comes to Global&Patch, that means the local information is preserved, and we gain a significant improvement. Note, the combination of global, patch, and frame achieves an accuracy of 57.1%. These results prove the merit of each representation. For the skip-fusion block, as summarised in Table 3b, the parallel spatial-temporal modelling is beneficial. As for the choice of cached features, the combination of features from the $\{4, 8, 12\}$_th layers of the visual encoder achieves the best result in accuracy. As for the interactive spatiotemporal reasoning, we introduced two interactions in this process, *i.e.* interaction between global and patch (Interactive-1), and interaction between patch and frame (Interactive-2), as shown in Fig. 3. As shown in Table 3c, dropping any of them results in performance degradation. In Table 3d, we explore the impact of different designs of text-guided temporal reasoning on performance. The results show that the combination of Q-Fusion and KV-Fusion, as shown in Eq. (3), leads to better performance, and it is beneficial to keep the text guidance for the support samples. As mentioned in Sec. 3.3, we use selection mechanisms in the fusion of the matching scores. The results reported in Table 3e demonstrate that the selective mechanism only works on text-visual matching. This is because the support-query matching is based on OTAM, which is a frame-level matching metric, so the reduction in the number of frames may lead to a performance compromise.

An Effectiveness Analysis. In order to demonstrate the effectiveness and efficiency of EMP-Net, we conduct comparisons from multiple perspectives with

Table 4: A comparison with the previous SOTA under different paradigms on SSv2-Small. The experimental results are obtained using GeForce RTX 3090, implemented by PyTorch under the same code framework.

(a) Comparison EMP-Net with existing paradigm on CLIP-FSAR with different frames.

Method	Backbone	Tune Param	Frame	Time/Iter		GPU Memory	Accuracy
				Train	Test		
CLIP-FSAR	Trainable CLIP	89M	2	0.46s	0.08s	15,821MiB	43.5%
			4	-	-	OOM	-
	Adapter CLIP	4M	2	0.46s	0.08s	12,423MiB	45.1%
			4	0.78s	0.14s	21,719MiB	51.6%
			8	-	-	OOM	-
	Frozen CLIP	3M	2	0.26s	0.08s	3,967MiB	44.2%
			4	0.42s	0.14s	4,897MiB	48.7%
			8	0.75s	0.25s	6,799MiB	50.1%
EMP-Net	Frozen CLIP	7M	2	0.34s	0.08s	4,365MiB	44.8%
			4	0.54s	0.14s	5,717MiB	52.2%
			8	1.02s	0.26s	8,565MiB	57.1%

(b) Comparison EMP-Net with existing paradigm on CLIP-FSAR with different shots and backbones.

Method	Shot	Encoder	Time/Iter		GPU Memory	Accuracy
			Train	Test		
EMP-Net	1	CLIP-ViT-B/16	1.02s	0.26s	8,565MiB	57.1%
	5		2.32s	0.68s	11,119MiB	65.7%
	10		3.94s	1.24s	15,795MiB	68.5%
	1	CLIP-ViT-B/32	0.92s	0.23s	4,293MiB	54.4%
		CLIP-ViT-B/16	1.02s	0.26s	8,565MiB	57.1%

existing SOTA, as summarised in Table 4. In Table 4a, we evaluate CLIP-FSAR under several different paradigms and compare those results with the proposed EMP-Net. We find that when using a trainable CLIP as the backbone for CLIP-FSAR, the training with more than 2 frames would cause the OOM error. When changing the training paradigm to adapter CLIP, the GPU memory compared with the trainable CLIP under the same setting reduces to a certain extent, but the saving is very limited. The GPU "out of memory" issue happens again when the input exceeds 8 frames under the adapter CLIP paradigm. Compared with those two paradigms, when we adopt the frozen CLIP, training cost drops a lot, but at the cost of a huge performance drop. The performance of CLIP-FSAR under frozen CLIP with 8 frames is even worse than that of adapter CLIP with 4 frames (50.1% *vs*. 51.6%). Compared with these paradigms, EMP-Net strikes a perfect balance between accuracy and the training overhead. EMP-Net only costs 8,565 MB GPU memory when taking 8 frames as input, which is significantly lower than CLIP-FSAR with the trainable or adapter CLIP. When comparing only the frozen CLIP as the backbone, EMP-Net achieves higher accuracy than CLIP-FSAR, while only requiring a very limited increase in the training overhead. As for the testing phase, EMP-Net basically does not increase the computing overhead.

We also evaluate the performance of EMP-Net under different settings. From Table 4b, as we demonstrated previously, EMP-Net achieves 57.1% under 1-shot evaluation. When it comes to 5 and 10 shots, our method further pushes the

Table 5: A comparison with the previous SOTA methods on temporal-related datasets. "*" represents the result implemented by [36].

Method	Reference	Pre-training	SSv2-Small		SSv2-Full	
			1-shot	5-shot	1-shot	5-shot
MatchingNet [35]	NeurIPS(16)	INet-RN50	31.3	45.5	-	-
MAML [9]	ICML(17)	INet-RN50	30.9	41.9	-	-
CMN++ [54]	ECCV(18)	INet-RN50	-	-	34.4	43.8
OTAM [3]	CVPR(20)	INet-RN50	36.4	48.0	42.8	52.3
ITANet [52]	IJCAI(21)	INet-RN50	39.8	53.7	49.2	62.3
TRX [25]	CVPR(21)	INet-RN50	36.0	56.7*	42.0	64.6
TA^{2}N [18]	AAAI(22)	INet-RN50	-	-	47.6	61.0
STRM [34]	CVPR(22)	INet-RN50	37.1	55.3	43.1	68.1
MTFAN [43]	CVPR(22)	INet-RN50	-	-	45.7	60.4
HyRSM [38]	CVPR(22)	INet-RN50	40.6	56.1	54.3	69.0
HCL [53]	ECCV(22)	INet-RN50	38.7	55.4	47.3	64.9
Huang *et al.* [13]	ECCV(22)	INet-RN50	38.9	61.6	49.3	66.7
Nguyen *et al.* [22]	ECCV(22)	INet-RN50	-	-	43.8	61.1
SloshNet [46]	AAAI(23)	INet-RN50	-	-	46.5	68.3
MoLo (OTAM) [37]	CVPR(23)	INet-RN50	41.9	56.2	55.0	69.6
CLIP-Freeze [27]	ICML(21)	CLIP-ViT-B/16	29.5	42.5	30.0	42.4
CapFSAR (OTAM) [39]	arXiv(23)	BLIP-ViT-B/16	45.9	59.9	51.9	68.2
CLIP-FSAR [36]	IJCV(23)	CLIP-ViT-B/16	54.6	61.8	62.1	72.1
D^{2}ST-Adapter [24]	arXiv(23)	CLIP-ViT-B/16	55.0	69.3	66.7	81.9
CLIP-CPM^{2}C [12]	arXiv(23)	CLIP-ViT-B/16	52.3	62.6	60.1	72.8
EMP-Net	Ours	CLIP-ViT-B/16	57.1	65.7	63.1	73.0

accuracy to 65.7% and 68.5%. For completeness, we also report the performance of EMP-Net using different ViT models as the encoder. From Table 4b, larger models lead to a better performance in accuracy. However, it is worth noting that when using ViT-B/32, EMP-Net can still achieve 54.4% in accuracy with much lower GPU memory, which is much better than other paradigms from Table 4a.

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

To validate the performance of our method, we chose multiple benchmarks to compare EMP-Net with the mainstream methods under 1-shot and 5-shot settings. The experimental results are summarised in Table 5 and Table 6. Especially, we compare our methods with several multi-modal pre-trained methods, including: trainable-based methods: CLIP-FSAR, CLIP-CPM^{2}C; adapter-based method: D^{2}ST-Adapter; frozen-based methods: CLIP-Freeze, CapFSAR.

On the temporal-related dataset, *i.e.* SSv2-Small and SSv2-Full, our method achieves significant improvement. From Table 5, we can see that: EMP-Net secures a huge lead when compared with the methods pre-trained on INet-RN50, which proves that our method takes advantage of the CLIP model; Given the same pre-trained model, EMP-Net achieves better performance compared with CLIP-FSAR; Compared to other frozen-based methods (CLIP-Freeze and CapFSAR), EMP-Net secures significant improvements, which proves that our method is more effective in feature utilisation under similar conditions. On the spatial-related datasets, *i.e.* HMDB-51, UCF-101, and K400, EMP-Net reaches the same level as other methods. The performance of our methods is slightly lower than other trainable and adapter based CLIP paradigms. However, it is worth noting that our method has a huge efficiency advantage over trainable and adapter

Table 6: A comparison with the previous SOTA methods on spatial-related datasets. "*" represents the result implemented by [36].

Method	Reference	Pre-training	HMDB-51		UCF-101		K400	
			1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MatchingNet [35]	NeurIPS(16)	INet-RN50	-	-	-	-	53.3	74.6
MAML [9]	ICML(17)	INet-RN50	-	-	-	-	54.2	75.3
ProtoNet [30]	NeurIPS(17)	-	54.2	68.4	74.0	89.6	64.5	77.9
TARN [2]	BMVC(19)	INet-RN50	-	-	-	-	64.8	78.5
ARN [51]	ECCV(20)	-	45.5	60.6	66.3	83.1	63.7	82.4
OTAM [3]	CVPR(20)	INet-RN50	54.5	68.0	79.9	88.9	72.2*	84.2*
ITANet [52]	IJCAI(21)	INet-RN50	-	-	-	-	73.6	84.3
TRX [25]	CVPR(21)	INet-RN50	53.1	75.6	78.2	96.1	63.6	85.9
TA^{2}N [18]	AAAI(22)	INet-RN50	59.7	73.9	81.9	95.1	72.8	85.8
STRM [34]	CVPR(22)	INet-RN50	52.3	77.3	80.5	96.9	62.9	86.7
MTFAN [43]	CVPR(22)	INet-RN50	59.0	74.6	84.8	95.1	74.6	87.4
HyRSM [38]	CVPR(22)	INet-RN50	60.3	76.0	83.9	94.7	73.7	86.1
HCL [53]	ECCV(22)	INet-RN50	59.1	76.3	82.5	93.9	73.7	85.8
Huang *et al.* [13]	ECCV(22)	INet-RN50	60.1	77.0	71.4	91.0	73.3	86.4
Nguyen *et al.* [22]	ECCV(22)	INet-RN50	59.6	76.9	84.9	95.9	74.3	87.4
SloshNet [46]	AAAI(23)	INet-RN50	59.4	77.5	86.0	97.1	70.4	87.0
MoLo (OTAM) [37]	CVPR(23)	INet-RN50	59.8	76.1	85.4	95.1	73.8	85.1
CLIP-Freeze [27]	ICML(21)	CLIP-ViT-B/16	58.2	77.0	89.7	95.7	78.9	91.9
CapFSAR (OTAM) [39]	arXiv(23)	BLIP-ViT-B/16	65.2	78.6	93.3	97.8	84.9	93.1
CLIP-FSAR [36]	IJCV(23)	CLIP-ViT-B/16	77.1	87.7	97.0	99.1	94.8	95.4
D^{2}ST-Adapter [24]	arXiv(23)	CLIP-ViT-B/16	77.1	88.2	96.4	99.1	89.3	95.5
CLIP-CPM^{2}C [12]	arXiv(23)	CLIP-ViT-B/16	75.9	88.0	95.0	98.6	91.0	95.5
EMP-Net	Ours	CLIP-ViT-B/16	76.8	85.8	94.3	98.2	89.1	93.5

methods, as shown on 4. Compared with other frozen-based methods, *i.e.* CLIP-Freeze and CapFSAR, our method still has obvious advantages.

Overall, compared with all other paradigms, our method achieves the best balance between accuracy and training overhead.

5 Conclusion and Limitations

To construct an effective solution for efficient FSAR, this paper presented EMP-Net, which combines multi-level representation and post-reasoning mechanism. EMP-Net includes the following innovations: Firstly, we proposed an effective domain transfer strategy, namely image2video adaption, which uses a completely frozen CLIP to provide rich prior knowledge. To expand the capacity and discriminability of the model, the reasoning and inference are performed in a multi-level manner. We also integrated text-to-visual matching and support-to-query matching on multi-level representations to enable a trustworthy inference. By combining the aforementioned designs, the proposed method significantly outperformed the latest frameworks in training overhead and accuracy.

Meanwhile, although our method achieves leading performance, there are still potential parts for improvement, including (1) *Scalability*: Our solution is based on Transformer and as such it is difficult to extend it to CNN-based structures. (2) *Flexibility*: EMP-Net does not consider the adaptive utilisation of different levels of information, which should theoretically yield higher results.

Acknowledgements

This work is supported in part by the National Key Research and Development Program of China (2023YFF1105102, 2023YFF1105105), the National Natural Science Foundation of China (62020106012, 62332008, 62106089, U1836218, 62336004), the 111 Project of Ministry of Education of China (B12018), and the UK EPSRC (EP/V002856/1, EP/T022205/1).

References

1. Ben-Ari, R., Nacson, M.S., Azulai, O., Barzelay, U., Rotman, D.: Taen: temporal aware embedding network for few-shot action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 2786–2794 (2021). https://doi.org/10.1109/CVPRW53098.2021.00313
2. Bishay, M., Zoumpourlis, G., Patras, I.: Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. arXiv preprint arXiv:1907.09021 (2019)
3. Cao, K., Ji, J., Cao, Z., Chang, C.Y., Niebles, J.C.: Few-shot video classification via temporal alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10618–10627 (2020). https://doi.org/10.1109/CVPR42600.2020.01063
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017). https://doi.org/10.1109/CVPR.2017.502
5. Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.C.F., Huang, J.B.: A closer look at few-shot classification. In: International Conference on Learning Representations (2019)
6. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part II 9. pp. 428–441. Springer (2006). https://doi.org/10.1007/11744047_33
7. Dhillon, G.S., Chaudhari, P., Ravichandran, A., Soatto, S.: A baseline for few-shot image classification. In: International Conference on Learning Representations (2020)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning. vol. 70, pp. 1126–1135. PMLR (2017)
10. Fu, M., Zhu, K., Wu, J.: Dtl: Disentangled transfer learning for visual recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 12082–12090 (2024). https://doi.org/10.1609/aaai.v38i11.29096
11. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5842–5850 (2017). https://doi.org/10.1109/ICCV.2017.622

12. Guo, F., Zhu, L., Wang, Y., Qi, H.: Consistency prototype module and motion compensation for few-shot action recognition (clip-cpm^2c). arXiv preprint arXiv:2312.01083 (2023)
13. Huang, Y., Yang, L., Sato, Y.: Compound prototype matching for few-shot action recognition. In: European Conference on Computer Vision. pp. 351–368. Springer (2022). https://doi.org/10.1007/978-3-031-19772-7_21
14. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 3192–3199 (2013). https://doi.org/10.1109/ICCV.2013.396
15. Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-shot object detection via feature reweighting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8420–8429 (2019). https://doi.org/10.1109/ICCV.2019.00851
16. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 International conference on computer vision. pp. 2556–2563. IEEE (2011). https://doi.org/10.1109/ICCV.2011.6126543
17. Lan, T., Zhu, Y., Zamir, A.R., Savarese, S.: Action recognition by hierarchical mid-level action elements. In: Proceedings of the IEEE international conference on computer vision. pp. 4552–4560 (2015). https://doi.org/10.1109/ICCV.2015.517
18. Li, S., Liu, H., Qian, R., Li, Y., See, J., Fei, M., Yu, X., Lin, W.: Ta2n: Two-stage action alignment network for few-shot action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1404–1411 (2022). https://doi.org/10.1609/aaai.v36i2.20029
19. Li, W., Wang, L., Xu, J., Huo, J., Gao, Y., Luo, J.: Revisiting local descriptor based image-to-class measure for few-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7260–7268 (2019). https://doi.org/10.1109/CVPR.2019.00743
20. Li, W., Wang, Z., Yang, X., Dong, C., Tian, P., Qin, T., Huo, J., Shi, Y., Wang, L., Gao, Y., et al.: Libfewshot: A comprehensive library for few-shot learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(12), 14938–14955 (2023). https://doi.org/10.1109/TPAMI.2023.3312125
21. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7083–7093 (2019). https://doi.org/10.1109/ICCV.2019.00718
22. Nguyen, K.D., Tran, Q.H., Nguyen, K., Hua, B.S., Nguyen, R.: Inductive and transductive few-shot video classification via appearance and temporal alignments. In: European Conference on Computer Vision. pp. 471–487. Springer (2022). https://doi.org/10.1007/978-3-031-20044-1_27
23. Pan, J., Lin, Z., Zhu, X., Shao, J., Li, H.: St-adapter: Parameter-efficient image-to-video transfer learning. Advances in Neural Information Processing Systems **35**, 26462–26477 (2022)
24. Pei, W., Tan, Q., Lu, G., Tian, J.: D^{2}st-adapter: Disentangled-and-deformable spatio-temporal adapter for few-shot action recognition. arXiv preprint arXiv:2312.01431 (2023)
25. Perrett, T., Masullo, A., Burghardt, T., Mirmehdi, M., Damen, D.: Temporal-relational crosstransformers for few-shot action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 475–484 (2021). https://doi.org/10.1109/CVPR46437.2021.00054

26. Qian, R., Lin, W., See, J., Li, D.: Controllable augmentations for video representation learning. Visual Intelligence **2**(1), 1 (2024). https://doi.org/10.1007/s44267-023-00034-7
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
28. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th ACM international conference on Multimedia. pp. 357–360 (2007). https://doi.org/10.1145/1291233.1291311
29. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems **27** (2014)
30. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in Neural Information Processing Systems **30** (2017)
31. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
32. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1199–1208 (2018). https://doi.org/10.1109/CVPR.2018.00131
33. Sung, Y.L., Cho, J., Bansal, M.: Lst: Ladder side-tuning for parameter and memory efficient transfer learning. Advances in Neural Information Processing Systems **35**, 12991–13005 (2022)
34. Thatipelli, A., Narayan, S., Khan, S., Anwer, R.M., Khan, F.S., Ghanem, B.: Spatio-temporal relation modeling for few-shot action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19958–19967 (2022). https://doi.org/10.1109/CVPR52688.2022.01933
35. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. Advances in Neural Information Processing Systems **29** (2016)
36. Wang, X., Zhang, S., Cen, J., Gao, C., Zhang, Y., Zhao, D., Sang, N.: Clip-guided prototype modulating for few-shot action recognition. International Journal of Computer Vision **132**(6), 1899–1912 (2024). https://doi.org/10.1007/s11263-023-01917-4
37. Wang, X., Zhang, S., Qing, Z., Gao, C., Zhang, Y., Zhao, D., Sang, N.: 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. pp. 18011–18021 (2023). https://doi.org/10.1109/CVPR52729.2023.01727
38. Wang, X., Zhang, S., Qing, Z., Tang, M., Zuo, Z., Gao, C., Jin, R., Sang, N.: Hybrid relation guided set matching for few-shot action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19948–19957 (2022). https://doi.org/10.1109/CVPR52688.2022.01932
39. Wang, X., Zhang, S., Yuan, H., Zhang, Y., Gao, C., Zhao, D., Sang, N.: Few-shot action recognition with captioning foundation models. arXiv preprint arXiv:2310.10125 (2023)
40. Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M.: Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys (csur) **53**(3), 1–34 (2020). https://doi.org/10.1145/3386252
41. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Computer Vision–ECCV 2008: 10th

European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part II 10. pp. 650–663. Springer (2008). https://doi.org/10.1007/978-3-540-88688-4_48
42. Wu, C., Wu, X.J., Kittler, J., Xu, T., Ahmed, S., Awais, M., Feng, Z.: Scd-net: Spatiotemporal clues disentanglement network for self-supervised skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5949–5957 (2024). https://doi.org/10.1609/aaai.v38i6.28409
43. Wu, J., Zhang, T., Zhang, Z., Wu, F., Zhang, Y.: Motion-modulated temporal fragment alignment network for few-shot action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9151–9160 (2022). https://doi.org/10.1109/CVPR52688.2022.00894
44. Xiang, W., Li, C., Wang, B., Wei, X., Hua, X.S., Zhang, L.: Spatiotemporal self-attention modeling with temporal patch shift for action recognition. In: European Conference on Computer Vision. pp. 627–644. Springer (2022). https://doi.org/10.1007/978-3-031-20062-5_36
45. Xing, J., Wang, M., Hou, X., Dai, G., Wang, J., Liu, Y.: Multimodal adaptation of clip for few-shot action recognition. arXiv preprint arXiv:2308.01532 (2023)
46. Xing, J., Wang, M., Liu, Y., Mu, B.: Revisiting the spatial and temporal modeling for few-shot action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3001–3009 (2023). https://doi.org/10.1609/aaai.v37i3.25403
47. Xing, J., Wang, M., Ruan, Y., Chen, B., Guo, Y., Mu, B., Dai, G., Wang, J., Liu, Y.: Boosting few-shot action recognition with graph-guided hybrid matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1740–1750 (2023). https://doi.org/10.1109/ICCV51070.2023.00167
48. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2945–2954 (2023). https://doi.org/10.1109/CVPR52729.2023.00288
49. Xu, W., Xu, Y., Wang, H., Tu, Z.: Attentional constellation nets for few-shot learning. In: International Conference on Learning Representations (2021)
50. Yang, Y., Cui, Z., Xu, J., Zhong, C., Zheng, W.S., Wang, R.: Continual learning with bayesian model based on a fixed pre-trained feature extractor. Visual Intelligence **1**(1), 5 (2023). https://doi.org/10.1007/s44267-023-00005-y
51. Zhang, H., Zhang, L., Qi, X., Li, H., Torr, P.H., Koniusz, P.: 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. pp. 525–542. Springer (2020). https://doi.org/10.1007/978-3-030-58558-7_31
52. Zhang, S., Zhou, J., He, X.: Learning implicit temporal alignment for few-shot video classification. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. pp. 1309–1315 (2021). https://doi.org/10.24963/ijcai.2021/181
53. Zheng, S., Chen, S., Jin, Q.: Few-shot action recognition with hierarchical matching and contrastive learning. In: European Conference on Computer Vision. pp. 297–313. Springer (2022). https://doi.org/10.1007/978-3-031-19772-7_18
54. Zhu, L., Yang, Y.: Compound memory networks for few-shot video classification. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 751–766 (2018). https://doi.org/10.1007/978-3-030-01234-2_46