Semantic-aware Video Representation for Few-shot Action Recognition

Yutao Tang
Johns Hopkins University
ytang67@jhu.edu

Benjamín Béjar
Paul Scherrer Institut
benjamin.bejar@psi.ch

René Vidal
University of Pennsylvania
vidalr@upenn.edu

Abstract

*Recent work on action recognition leverages 3D features and textual information to achieve state-of-the-art performance. However, most of the current few-shot action recognition methods still rely on 2D frame-level representations, often require additional components to model temporal relations, and employ complex distance functions to achieve accurate alignment of these representations. In addition, existing methods struggle to effectively integrate textual semantics, some resorting to concatenation or addition of textual and visual features, and some using text merely as an additional supervision without truly achieving feature fusion and information transfer from different modalities. In this work, we propose a simple yet effective **S**emantic-**A**ware **F**ew-**S**hot **A**ction **R**ecognition (**SAFSAR**) model to address these issues. We show that directly leveraging a 3D feature extractor combined with an effective feature-fusion scheme, and a simple cosine similarity for classification can yield better performance without the need of extra components for temporal modeling or complex distance functions. We introduce an innovative scheme to encode the textual semantics into the video representation which adaptively fuses features from text and video, and encourages the visual encoder to extract more semantically consistent features. In this scheme, SAFSAR achieves alignment and fusion in a compact way. Experiments on five challenging few-shot action recognition benchmarks under various settings demonstrate that the proposed SAFSAR model significantly improves the state-of-the-art performance.*

1. Introduction

Few-shot action recognition (FSAR) is a challenging and practical problem in the field of computer vision. Unlike conventional action recognition which typically relies on large amounts of annotated data, and where the goal at inference time is to correctly classify actions that have been seen during training, FSAR aims to develop models capable of accurately identifying unseen action categories through just a handful of annotated samples per category. Consequently, FSAR is a promising alternative to conventional action recognition when collecting abundant annotated data for each action category is impractical or very costly [61]. Specifically, an FSAR problem involves a support set comprised of a few labelled videos and a corresponding query set which contains unlabelled videos. The goal of FSAR is to assign a correct label to query videos merely using those few labelled videos in the support set as reference.

Mainstream approaches to FSAR [5, 6, 34, 51, 56, 57, 59, 61] rely on metric-based meta-learning [38]. Such approaches first learn to generate representations for both support and query videos, and then learn a similarity metric to associate query videos with one of the categories in the support set, akin to the nearest neighbor classifier. For example, OTAM [6] extracts frame-level representations and designs an ordered temporal alignment distance function to measure the similarity between support and query videos. TRX [34] exploits the CrossTransformer [14] to highlight query-related sub-sequences of support videos and constructs tuples of varying number of frames to compare support and query videos. HyRSM [51] proposes a hybrid relation module that explores within- and cross-video relations to extract task-specific representations and designs a set matching metric to get the query-support correspondence.

Despite recent advances in the field, current approaches still have two considerable limitations. First, most of the state-of-the-art (SoTA) methods rely on 2D feature extractors to obtain frame-level representations, employ additional components to model temporal dependencies and enforce temporal alignment with a complex distance function. While this type of approaches achieve considerable classification results, the advances in general video action recognition indicate that 3D feature extractors, such as 3D-CNNs [7, 15, 20] and video transformers [3, 4, 42], hold significant promise in capturing spatio-temporal features more accurately. This is particularly evident during action sequences characterized by intricate temporal movements, where these advanced models demonstrate their potential to excel. In light of these advances, it is beneficial to leverage 3D feature extractors, avoiding the need of extra components or intricate distance functions in the FSAR task. Additionally, the second important limitation of current approaches for FSAR concerns the integration and fusion of textual information with visual cues. Previous attempts struggle to fully leverage the textual descriptors due to simplistic fusion strategies [49, 58] or by treating text as merely an additional supervision without truly fusing features from different modalities [24, 37]. For example, [58] employs an LSTM-alike memory network to get dynamic visual and textual features but then simply concatenate them as multimodal features for few-shot recognition. [49] learns a generative model to project visual features into the textual semantic space, and sums the visual features and the generated semantic features element wise for few-shot classification. [32] fuses visual and textual features through weighted average. [24] treats text as a supplemental supervision to regularize the visual features of the query video without truly fusing the textual features with the visual features. Besides, these methods do not harness the full potential of the pre-trained language models as they use Word2Vec [30] embeddings or handcrafted sentence templates, *e.g. a video of* $\{$*action*$\}$. These shortcomings substantially hinder effective integration of textual semantics.

In this paper, we address the aforementioned two limitations by proposing a novel **S**emantic-**A**ware **F**ew-**S**hot **A**ction **R**ecognition (**SAFSAR**) model. Specifically, we leverage recent advances in general action recognition and use a 3D feature extractor to directly extract a discriminative spatio-temporal representation for both support and query videos. Further, we design a novel paradigm that achieves seamless integration of the textual semantics by aligning and fusing the visual and textual features in a compact way. Specifically, we first develop a lightweight fusion module to adaptively encode textual semantics into the video representations of the support set while the query video representations remain untouched. Next, we apply a task-specific learning module to enrich the video representations with the interactions across the support and query videos. Finally, we classify the query video using a cosine similarity distance between the final representations of the support and query videos. This paradigm design provides a natural mechanism for textual semantics and visual features to be learned and fused in a shared latent knowledge space because we are aligning semantic-unaware query representations with semantic-aware support representations. By this means, we not only effectively fuse features from the two modalities but also encourage the visual encoder to extract more semantically consistent features.

In summary, our contributions are three-fold. (1) We propose a simple yet effective multi-modal model (SAFSAR) for few-shot action recognition that extracts discriminative spatio-temporal features while simultaneously exploiting textual semantics. (2) We design a novel paradigm for aligning and fusing visual features and textual semantics in a compact way. (3) We perform extensive evaluations on five challenging benchmarks to demonstrate the superiority of our approach over prior art.

2. Related Work

Few-shot Learning. Few-shot learning approaches can be broadly divided into three categories: data-level, optimization-level, and metric-level approaches. The data-level approaches [10, 13, 35, 47] address the few-shot learning problem by enlarging the datasets through collection or synthesis of data to enhance model generalization for tasks with scarce data. They accomplish this using pretraining and fine-tuning techniques. The optimization-level approaches [1,2,16,33,36,55] learn an optimized model that can serve as a good initialization and that can be quickly adapted to a new task given only limited amount of data. The metric-level approaches concentrate on developing a latent feature embedding space where the similarities between the query set and the support set representations can be measured effectively. For example, Matching Network [46] is among the first to implement the latter idea for one-shot learning. ProtoNet [38] extends it and proposes to use the average embeddings as the prototype of each class in the few-shot setting and uses the Euclidean distance instead of cosine similarity to compare the embeddings. Relation Network [40] further improves it by employing a learned distance function. Our work is a metric-level approach that shares the same spirit of “learn-to-compare” and we focus on the more challenging few-shot action recognition task which involves diverse spatio-temporal dependencies.

Unimodal FSAR. Few-shot action recognition (FSAR) is a sub-field of few-shot learning and it deals with complex high-dimensional videos. Existing techniques for FSAR generally follow the metric-level approaches, typically using a single modality, *i.e*. the RGB videos alone. In terms of feature extraction, most methods use 2D feature extractors, like ResNet-50 [21], to extract frame-level representations. CMN [61] uses a multi-saliency embedding algorithm to selectively combine the frame-level representations into a video representation. OTAM [6] compares the support and query videos using an ordered temporal alignment distance function on the frame-level representations. ITANet [59] proposes a frame-wise implicit temporal alignment network to model the temporal context from the frame-level representations. TRX [34] exploits CrossTransformer [14] to highlight the query-related sub-sequences of support videos and matches videos through plentiful tuples of varying number of frames. STRM [41] applies spatio-temporal enrichment to the frame-level representations and uses TRX with a single cardinality to classify the query video. HyRSM [51] proposes a hybrid relation module that models within- and cross-video relations to extract task-

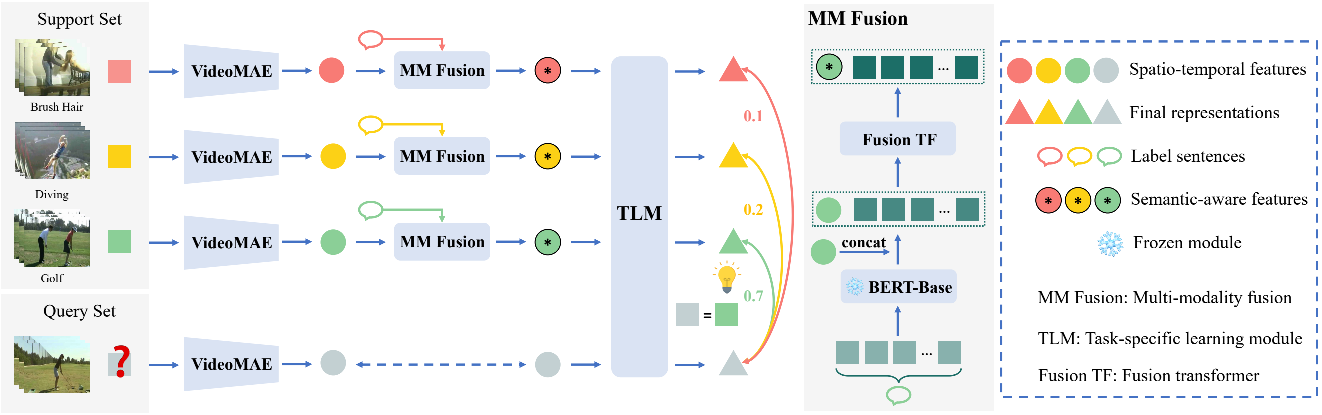


Figure 1. **Left:** Model architecture of our proposed SAFSAR illustrated in a 3-way 1-shot problem. We first extract spatio-temporal features from the support and query videos using VideoMAE. Then, the support video features and their corresponding label sentences are input to the multi-modality fusion (MM Fusion) module to obtain semantic-aware support features, while the query video features remain untouched. Finally, the support and query features are input to a task-specific learning module (TLM) to get the final representations which are then used to compute the cosine similarity scores between the query and support videos. Subsequently, we apply softmax to the cosine similarity scores, obtaining the probabilities of classifying the query video against support action classes. **Middle:** Details of the MM Fusion module. The label description sentence is first tokenized and then forwarded through the frozen BERT-Base model, obtaining textual features for each token. Then, we concatenate the support features as an extra token to the textual feature tokens, which are then forwarded through a fusion transformer (Fusion TF) to obtain the semantic-aware support features. **Right:** The legend of this figure.

specific representations and a bidirectional Mean Hausdorff Metric (Bi-MHM) to get the query-support correspondence. Besides, very few approaches attempt at incorporating 3D feature extractors to directly extract spatio-temporal representations. TARN [5] uses C3D [43] to extract snippet features and apply snippet-based attention to achieve temporal alignment with a deep-distance metric. ARN [57] also uses C3D [43] for capturing short-range temporal context and introduces permutation-invariant pooling with spatial and temporal attention to obtain robust descriptors for videos. However, these approaches have inferior performance to recent methods using 2D features [41, 51]. Moreover, the advances on the video transformers [3, 4, 29, 42] suggest its promising application in FSAR. MASTAF [56] proposes a model-agnostic network with self- and cross-attention to classify the query video. They demonstrate using ViViT [3] achieves the best results compared to using 2D- or 3D-CNNs. However, they only explore 1-shot setting using ViViT. Our work leverages VideoMAE [42], a Transformer-based [45] model that is capable of capturing intricate spatio-temporal relationships within videos. We explore 1-shot, 3-shot, and 5-shot settings and achieve superior performance than MASTAF [56] as shown in Tab. 1.

Multi-modal FSAR. Some recent work has explored the inclusion of additional modalities such as motion [50, 53, 54], depth [17], object features [23] and text [24, 32, 37, 48, 49, 58]. Of particular interest are those approaches that explore the use of textual information in conjunction with visual cues for improved FSAR. [58] proposes a memory network to extract visual and textual features dynamically and then fuses them by concatenation for further recognition. [49] trains a generative model to imitate the semantic labels and fuses the visual features and the generated textual features by element-wise addition. [37] leverages the pre-trained vision-language model to generate matching scores between video frames and handcrafted action-related phrases. These scores are subsequently used as descriptors for each video to train a temporal network for action classification. [32] computes prototypes for both modalities respectively and then fuses them through a weighted average. [48] obtains enhanced visual features by channel-wise multiplication of the visual vectors and the semantic vectors. [24] uses semantic labels as a supplemental supervision to regularize the visual features of the query video without truly fusing the textual features with the visual features. While promising progress has been made, these approaches fall short of fully integrating textual and visual features because either they resort to simplistic fusion strategies, or they simply treat text as an auxiliary source of guidance instead of truly integrating it with visual features. This gap in the literature inspired us to design a novel paradigm in SAFSAR to combine multi-modal features seamlessly.

3. Methods

3.1. Few-Shot Action Recognition

We consider the following setup for multi-modal FSAR where we have access to a training set of triplets $\{(y_i, \boldsymbol{x}_i, \tau_i)\}_{i=1}^{n}$, where $y_i \in \{1, \ldots, C\}$ is the action class label, $\boldsymbol{x}_i$ represents a video clip of T_i frames, and τ_i is a text description of the activity. In the few-shot learning paradigm, a subset of labeled videos, *i.e*. the *support set* is used to classify a novel unlabeled *query* video. We follow the N-way K-shot approach which means that the support set contains a subset of videos that represent $N \leq C$ of the total number of classes, and with K instances per class. The query video has an unknown class label belonging to one of the classes in the support set. To make the training more consistent with the testing scenario, we adopt episodic training as in [6, 34, 51]. Specifically, at each training episode $e \in 1, \ldots, E$, where E is the number of episodes, we extract a support set $\mathcal{S}_e \subset \{1, \ldots, n\}$ as follows: We first randomly sample a subset of N classes out of the total number of classes C and for each class we sample K instances at random without replacement. This means that the support set contains a total of $N \times K$ triplets from the training set. Next, we sample a query set $\mathcal{Q} = \{q_e\}$ consisting of a single video from one of the classes in the support set, where q_e is sampled at random from $\{1, \ldots, n\} \backslash \mathcal{S}_e$. A classifier is then trained using all episodes.

3.2. The SAFSAR model

The overall architecture of the SAFSAR model is shown in Fig. 1. We first sample T frames uniformly from each video to extract spatio-temporal features using VideoMAE [42]. Then, the support video features and their corresponding label sentences are processed by the multi-modality fusion (MM Fusion) module to obtain semantic-aware support features, while query video features remain untouched. These features are subsequently input to a task-specific learning module (TLM) to reinforce the interactive cues across videos, getting the final representations. Finally, a cosine similarity distance is computed between query and support final representations to classify the query video.

3.2.1 Spatio-temporal Feature Extraction

Common approaches to FSAR rely on per-frame (2D) feature extraction and later modeling of the temporal dependencies of the extracted spatial features. For instance, in HyRSM [51], they first extract features for each frame and then design a self-attention module operating on the frame-level features to capture long-range temporal dependencies. Arguably, such representations might be suboptimal in capturing the relevant spatio-temporal features for the classification task. In our SAFSAR approach, we propose instead to rely directly on spatio-temporal representations extracted using VideoMAE [42]. If we denote $\Phi_{VMAE}(\cdot)$ as the feature extractor for VideoMAE, we can compute video-based (3D) representations for an input video $\boldsymbol{x}$ to obtain a d-dimensional spatio-temporal features:

$$\boldsymbol{f} = \Phi_{VMAE}(\boldsymbol{x}), \quad \boldsymbol{f} \in \mathbb{R}^d. \tag{1}$$

Then, we follow ProtoNet [38] and use averaged video features in the same category as the class prototypes for the support set. Specifically, for a given episode with support set $\mathcal{S}_e$ we compute:

$$\bar{\boldsymbol{f}}_c = \frac{1}{K} \sum_{j \in \mathcal{S}_e} \mathbb{1}_{(y_j = c)} \Phi_{VMAE}(\boldsymbol{x}_j), \quad c \in \mathcal{C}_e, \tag{2}$$

where $\mathcal{C}_e \subseteq \{1, \ldots, C\}$ is the subset of classes represented in the support set $\mathcal{S}_e$, and where $\mathbb{1}_{(\cdot)}$ is a $0/1$ indicator function taking the value 1 if the argument is true. These class prototypes $\bar{\boldsymbol{f}}_c$ are then input to the following modules.

3.2.2 Multi-modality Fusion Module

In this module, the objective is to integrate textual semantics into the visual features. The previous attempts mainly extract embeddings of the label keywords [24, 48, 49, 49], *e.g*., `diving`, `brush hair`, `golf`, *etc*. using Word2Vec [30] or utilize handcrafted sentence templates [32, 37], *e.g*., *a video of* $\{$*action*$\}$, as input to a pre-trained text encoder to extract textual semantics. However, employing these strategies does not exploit the full capacity of the pre-trained language models. In our work, we use more detailed and elaborated descriptions provided by [9] to better capture nuanced aspects of actions, and enable more meaningful alignment and fusion of both modalities. The elaborated description in [9] extends from the class names to a comprehensive definition of the action classes. For example, `Archery` can be expanded as *shooting with a bow and arrows, especially at a target as a sport*, and `LongJump` can be extended as *an athletic event in which competitors jump as far as possible along the ground in one leap*. With the inclusion of these descriptive phrases of the actions, we tap into the advantages of advanced language models to extract more enhanced and discriminative textual semantics.

As shown in the middle part of the Fig. 1, we first follow the rules in BERT [12] to tokenize the label description sentence into word tokens $\tau \in \mathbb{R}^L$, where L is the number of words in the description sentence. Then, we pass the word tokens through the frozen pre-trained BERT-Base model to get the textual features $\boldsymbol{s} \in \mathbb{R}^{L \times d}$ where d is the dimension of the textual features. Next, we concatenate the class prototype $\bar{\boldsymbol{f}}_c$ as an extra token to the textual features. The concatenation is then input to the Fusion Transformer module (Fusion TF) which consists of ℓ layers of Transformer [45]. The semantic-aware visual features are then computed as:

$$[\boldsymbol{f}_c^*, \boldsymbol{s}^*] = \Phi_{FTF}([\bar{\boldsymbol{f}}_c, \boldsymbol{s}]), \quad \boldsymbol{s} = \Phi_{BERT}(\tau), \tag{3}$$

where $\Phi_{BERT}(\cdot)$ refers to the frozen BERT-Base model, $[\bar{\boldsymbol{f}}_c, \boldsymbol{s}]$ represents the concatenation of the class prototype and the textual features, $\Phi_{FTF}(\cdot)$ denotes the Fusion Transformer module, which adaptively integrates the textual semantics into the visual features, $\boldsymbol{s}^*$ is the output sentence

Methods	SSv2-Full			SSv2-Small			UCF101			HMDB51		
	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot
Unimodal FSAR												
ARN [57]	-	-	-	-	-	-	66.30	-	83.10	45.50	-	60.60
OTAM [6]	42.80	51.50	52.30	36.40	45.90	48.00	79.90	87.00	88.90	54.50	65.70	68.00
CMN-J [62]	-	-	-	36.20	44.60	48.80	-	-	-	-	-	-
ITANet [59]	49.20	59.10	62.30	39.80	49.40	53.70	-	-	-	-	-	-
TRX (Ω=1) [34]	38.80	54.40	60.60	34.90		47.60	53.30	-	-	-	-	-
TRX (Ω=2,3) [34]	42.00	57.60	64.60	36.00	51.90	59.10	78.20	92.40	96.10	53.10	66.80	75.60
TA^{2}N [28]	47.60	-	61.00	-	-	-	81.90	-	95.10	59.70	-	73.90
STRM [41]	-	-	70.20	-	-	-	-	-	98.10	-	-	81.30
MTFAN [54]	45.70	-	60.40	-	-	-	84.80	-	95.10	59.00	-	74.60
HyRSM [51]	54.30	65.10	69.00	40.60	52.30	56.10	83.90	93.00	94.70	60.30	71.70	76.00
TRX+L2A [18]	-	-	-	-	-	-	79.20	93.20	96.30	51.90	68.20	77.00
ATA [31]	43.80	-	61.10	-	-	-	84.90	-	95.90	59.60	-	76.90
HCL [60]	47.30	59.00	64.90	38.70	49.10	55.40	82.50	91.00	93.90	59.10	71.20	76.30
CPM [23]	49.30	-	66.70	38.90	-	61.60	71.40	-	91.00	60.10	-	77.00
MASTAF (TSN) [56]	46.90	-	62.40	37.50	-	50.20	79.30	-	90.3	54.80	-	67.70
MASTAF (R3D) [56]	50.30	-	66.70	39.90	-	52.20	90.60	-	97.60	67.90	-	81.20
MASTAF (ViViT) [56]	60.70	-	-	45.60	-	-	91.60	-	-	69.50	-	-
Multi-modal FSAR												
CMMN [58]	-	-	-	-	-	-	78.92	-	87.72	58.92	-	70.45
SRPN [49]	-	-	-	-	-	-	86.50	93.80	95.80	61.60	72.50	76.20
KP [37]	-	-	-	-	-	62.40	-	-	99.40	-	-	**87.40**
MORN [32]	-	-	71.70	-	-	-	-	-	97.70	-	-	87.10
TADRNet [48]	43.00	61.10	70.20	-	-	-	86.70	94.30	96.40	64.30	74.50	78.20
EANT [24]	-	-	-	-	-	-	87.00	94.60	96.20	62.50	73.10	77.20
SAFSAR (ours)	**71.26**	**74.93**	**76.75**	**59.23**	**61.49**	**63.68**	**98.30**	**99.31**	**99.54**	**77.38**	**83.74**	85.63

Table 1. Comparison to the state-of-the-art FSAR methods on the meta-testing set of SSv2-Full, SSv2-Small, UCF101, and HMDB51. The experiments are conducted under 5-way 1-shot, 3-shot, and 5-shot settings. "-" indicates the result is not available for this setting in that paper. Numbers in bold refer to the best results and numbers underlined refer to the second best results.

Methods	Epic-kitchens		
	1-shot	3-shot	5-shot
OTAM	46.00	53.90	56.30
TRX	43.40	53.50	58.90
HyRSM	47.40	56.40	**59.80**
SAFSAR (ours)	**54.02**	**58.03**	59.31

Table 2. Results on EPIC-KITCHENS under 5-way 1-shot, 3-shot, and 5-shot settings.

features of the Multi-modal Fusion module and $\boldsymbol{f}_c^*$ denotes the semantic-aware support features for class c. This paradigm provides a natural mechanism for encouraging the visual encoder to extract semantically consistent features as the goal is to match the semantic-unaware query video features to semantic-aware support video representations and also facilitate inter-modal feature integration to create coherent representations.

3.2.3 Task-specific Learning Module

Learning the features of each individual video independently can result into suboptimal performance in FSAR. As evidenced in previous studies, such an approach is often prone to overfitting to irrelevant information and might deteriorate the generalizability of the model to unseen classes [22, 27]. To address these concerns, we introduce the Task-specific Learning module (TLM), comprised of one Transformer [45] layer, whose purpose is to augment the video features via cross-video interactions, in order to yield adapted and discriminative representations tailored towards solving the current classification task at each episode. More precisely, we take the semantic-aware support representations and the untouched query features as tokens and apply the TLM on top of them to retrieve the final representations. We denote it as

$$[\{\tilde{\boldsymbol{f}}_c\}_{c\in\mathcal{C}_e}, \tilde{\boldsymbol{f}}_{q_e}] = \Phi_{TLM}\left([\{\boldsymbol{f}_c^*\}_{c\in\mathcal{C}_e}, \boldsymbol{f}_{q_e}]\right) \tag{4}$$

where $\tilde{\boldsymbol{f}}_c$, $c \in \mathcal{C}_e$ and $\boldsymbol{f}_{q_e}$ represent the final representations of the support videos and the query video, respectively, and

$\Phi_{TLM}(\cdot)$ denotes the Task-specific Learning module.

3.2.4 The Distance Function

Unlike other methods based on 2D feature extractors that rely on the design of complex distance functions for temporal alignment, and since our approach is inherently 3D, we use instead the cosine similarity to compute the distance between the final representations of the query video and those of the support videos. The cosine similarity between two feature vectors $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ is defined as:

$$\cos(\boldsymbol{f}_1, \boldsymbol{f}_2) = \frac{\langle \boldsymbol{f}_1, \boldsymbol{f}_2 \rangle}{\|\boldsymbol{f}_1\|_2 \|\boldsymbol{f}_2\|_2}, \tag{5}$$

where $\langle \cdot, \cdot \rangle$ denotes inner product. Then, we apply softmax to the cosine similarity scores, resulting in the probability of predicting the query video q_e to be a certain action class c. Specifically,

$$p_c = \frac{\exp(\cos(\tilde{\boldsymbol{f}}_c, \tilde{\boldsymbol{f}}_{q_e}))}{\sum_{j \in \mathcal{C}_e} \exp(\cos(\tilde{\boldsymbol{f}}_j, \tilde{\boldsymbol{f}}_{q_e}))}. \tag{6}$$

3.3. The Training Losses

We use the cross-entropy loss to train the classifier for the correct class assignment:

$$L_1 = - \sum_{c \in \mathcal{C}_e} \mathbb{1}_{(y_{q_e} = c)} \log p_c. \tag{7}$$

Additionally, in order to regularize the representations and improve generalization, we include a global classification loss into the training. Specifically, we pass the support and query video features extracted by VideoMAE through one linear layer $\boldsymbol{W} \in \mathbb{R}^{C \times d}$ and then apply a cross-entropy loss to supervise the classification against the total action categories C. We denote this loss as

$$\begin{aligned} L_2 = -\frac{1}{NK} \sum_{j \in \mathcal{S}_e} \sum_{c=1}^{C} \mathbb{1}_{(y_j = c)} \log \sigma_c\big(\boldsymbol{W} \boldsymbol{f}_j\big) \\ - \sum_{c=1}^{C} \mathbb{1}_{(y_{q_e} = c)} \log \sigma_c\big(\boldsymbol{W} \boldsymbol{f}_{q_e}\big) \end{aligned} \tag{8}$$

where we use $\sigma_c(\boldsymbol{v})$ to denote the c-th component of the softmax operation on vector $\boldsymbol{v}$ (*i.e.*, $\sigma_c(\boldsymbol{v}) = \exp(v_c) / \sum_j \exp(v_j)$).

Finally, the total loss for training SAFSAR can be expressed as the weighted sum of the cosine similarity classification loss and the global classification loss,

$$L = L_1 + \lambda L_2 \tag{9}$$

where λ is a weighting factor.

4. Experiments

Datasets. We evaluate our method on five few-shot action recognition benchmarks, namely HMDB51 [26], SSv2-Full [19], SSv2-Small [19], UCF101 [39], and EPIC-KITCHENS [11]. We do not evaluate on Kinetics-100 [7] because our feature extractor was pre-trained on the Kinetics dataset. We adopt the few-shot splits from OTAM [6] and CMN [61] for SSv2-Full and SSv2-Small, respectively, where SSv2-Full contains 10× more videos per action category in the training set. They both consist of 64, 12, and 24 classes for training, validation, and testing. For HMDB51 and UCF101, we use the few-shot splits from ARN [57]. HMDB51 contains 31, 10, and 10 classes while UCF101 contains 70, 10, and 21 classes for training, validation, and testing, respectively. EPIC-KITCHENS is a large-scale egocentric video dataset that includes various actions in kitchens. We use the splits from HyRSM [51] which contains 60 and 20 classes for training and testing.

Implementation Details. Experiments are conducted using the 5-way K-shot ($K \in \{1, 3, 5\}$) settings. We use VideoMAE-Base [42] as the feature extractor which is initialized with Kinectis-400 pre-trained weights from [42] for SSv2-Full and SSv2-Small. For all the other benchmarks, we initialize it with Kinetics-710 pre-trained weights from [52]. We uniformly sample $T = 8$ frames as in previous methods for a fair comparison. For the text encoder, we utilize BERT-Base [12] initialized with weights provided by [8]. In terms of the label descriptive sentences, we use the elaborate descriptions in [9] for HMDB51 and UCF101. For SSv2, we directly use the class names as they already provide detailed descriptions for each action class. For EPIC-KITCHENS, we use OpenAI's ChatGPT-3.5 to generate elaborative descriptions based on the class names. During training, we only freeze the patch embedding layer of VideoMAE-Base and the entire 12 layers of BERT-Base and use Adam [25] optimizer to train the model. Basic data augmentation like random cropping, flipping, and color jittering is applied. During inference, we report average accuracy over $E = 10,000$ episodes randomly sampled from the test set.

4.1. Comparison with State-of-the-art Methods

In this section, we evaluate the efficacy of our proposed SAFSAR by comparing its performance against existing state-of-the-art (SoTA) approaches on several challenging few-shot action recognition benchmarks. As outlined in Tab. 1, we order the methods by the publication year and categorize the SoTA methods into two groups, unimodal FSAR and multi-modal FSAR where the latter group uses text as additional information in their models. Remarkably, most research efforts in the multi-modal FSAR group focus on UCF101 and HMDB51, whereas fewer studies examine SSv2. Note that SSv2 is a different type of dataset

Number of Frames T	SSv2-Full		SSv2-Small		UCF101		HMDB51	
	1-shot	3-shot	1-shot	3-shot	1-shot	3-shot	1-shot	3-shot
8 frames	71.26	74.93	59.23	61.49	98.30	99.31	77.38	83.74
16 frames	**74.66**	**78.72**	**60.69**	**64.88**	**99.23**	**99.57**	**78.38**	**85.71**

Table 3. Comparison between the performance of sampling 8 frames and 16 frames under 5-way 1-shot and 3-shot settings.

as it is primarily a motion-centric dataset which emphasizes temporal modeling of the actions while UCF101 and HMDB51 are scene-centric datasets which tend to stress on scene understanding. Moreover, EPIC-KITCHENS is another challenging dataset due to its motion-centric nature and videos captured from a first-person perspective. In addition, this dataset exclusively involves actions performed within a kitchen setting, encompassing a lot of fine-grained classes that demand a comprehensive understanding of the actions. In our experiments, we not only evaluate our proposed SAFSAR in commonly used UCF101 and HMDB51, but also validate it in SSv2 and EPIC-KITCHENS. Our exceptional results verify the effectiveness and robustness of our approach in various types of datasets.

As shown in Tab. 1 and Tab. 2, we observe that our proposed SAFSAR achieves remarkable performance and consistently exhibits substantial advantages over other SoTA methods, especially under the challenging 1-shot and 3-shot settings. Concretely, SAFSAR provides around 11%, 14%, 7%, 8%, and 7% absolute improvement in SSv2-Full, SSv2-Small, UCF101, HMDB51, and EPIC-KITCHENS respectively, under the most challenging 5-way 1-shot scenario. Additionally, SAFSAR also unequivocally outpaces the other methods by a large margin under 3-shot setting. Turning to the 5-shot scenario, SAFSAR continues to outperform the other methods across the majority of benchmarks. While on HMDB51 and EPIC-KITCHEN, SAFSAR's performance is slightly below the state-of-the-art, it still maintains a competitive edge on SSv2-Full, delivering a noteworthy 5% improvement. Of specical note, our SAFSAR achieves substantial improvements on SSv2-Full and SSv2-Small benchmarks, which provides compelling evidence for the superiority of our approach in modeling temporal relations of the actions.

4.2. Analysis on the Properties of SAFSAR

Number of Frames. In this section, we conduct a thorough investigation of the effect of sampling different number of frames for SSv2-Full, SSv2-Small, UCF101, and HMDB51 under 5-way 1-shot and 3-shot settings. Throughout these experiments, the only change is that we freeze the first 6 layers along with the patch embedding layer and only finetune the last 6 layers in VideoMAE. Motivated by the fact that VideoMAE was pre-trained using 16 frames per video, we may reduce computation cost by fixing early layers, making it comparable with the computation cost of sampling 8 frames. The results in Tab. 3 reveal that sampling 16 frames is consistently better than sampling 8 frames. Remarkably, compared with the results of sampling 8 frames, sampling 16 frames brings about 4%, 3%, 0.26%, and 2% absolute improvement in SSv2-Full, SSv2-Small, UCF101, and HMDB51, respectively, under the challenging 5-way 3-shot scenario. These positive effects also demonstrate in the 1-shot setting, where an average increase of approximately 1% can be seen across all four benchmarks. The superiority of sampling 16 frames suggests that our proposed SAFSAR is capable of leveraging additional sampled frames to derive richer and better discriminative features, surpassing previous methods. For instance, we notice that although employing increased number of frames generally leads to higher recognition accuracy, some algorithms, such as HyRSM [51], exhibit saturated performance beyond 8 frames as revealed in their ablation analysis. Our findings imply that SAFSAR is capable of making full use of extra frames to achieve superior performance, confirming its strength in accurately encoding temporal relations.

Number of Layers l	SSv2-Small		HMDB51	
	1-shot	3-shot	1-shot	3-shot
1	57.48	61.38	77.33	83.36
2	**59.23**	**61.49**	**77.38**	**83.74**
3	56.73	60.83	77.31	82.69

Table 4. Results of using different l in Fusion FT module under 5-way 1-shot and 3-shot settings.

Number of Layers in the Fusion TF Module. As described in Sec. 3.2, the Fusion TF module consists of l layers of Transformer [45]. In this section, we study the impact of varying l on the performance of SAFSAR in the SSv2-Small and HMDB51 benchmarks, considering both 5-way 1-shot and 3-shot scenarios. As observed in Tab. 4, it appears that setting $l = 2$ yields the best scores under both scenarios across both benchmarks. Therefore, we opt to set $l = 2$ as the default setting in all other experiments unless otherwise noted.

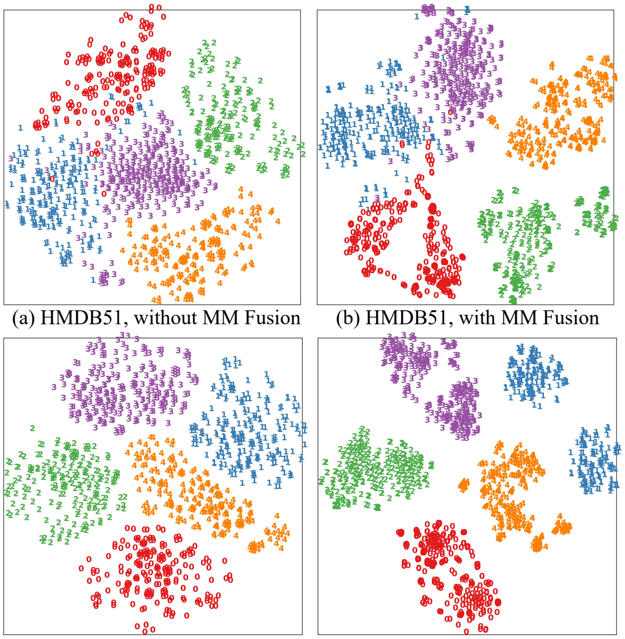
Ablation of the Proposed Modules. We summarize the effects of incorporating MM Fusion module and TLM in our proposed SAFSAR in Tab. 5. The first row serves as a baseline where neither of the module is used. We conduct the experiments in SSv2-Small and SSv2-Full under 5-way 1-shot and 3-shot settings. Particularly, we notice significant improvements when using both modules together, achieving on average about 15% increase over the baseline for 1-shot setting. Even only using the MM Fusion module alone, it brings appreciable 4% and 9% improvement for SSv2-Small and SSv2-Full, respectively, under 1-shot scenario. Similar trend can be observed for the 3-shot setting. Overall, the results confirm the effectiveness of our proposed modules.

MM Fusion module	TLM	SSv2-Small		SSv2-Full	
		1-shot	3-shot	1-shot	3-shot
		37.87	46.00	61.18	70.96
	✓	38.88	49.20	59.30	66.51
✓		41.14	48.19	70.51	73.94
✓	✓	**59.23**	**61.49**	**71.26**	**74.93**

Table 5. Ablation study on incorporating MM Fusion module and TLM under 5-way 1-shot and 3-shot settings.

Analysis of Computational Efficiency. We run one episode under 5-way 1-shot setting on one NVIDIA RTX A5000 GPU and present the computational efficiency analysis in Tab. 6. We observe that SAFSAR has more parameters compared to the other methods. The increase is attributed to our utilization of a 3D feature extractor and the incorporation of a language model which constitutes more than half of the total parameter count. Consequently, this architecture choice results in an elevated computational load. While our model does require more parameters and computations, this investment is rewarded by substantial performance gains. For example, SAFSAR demonstrates an average increase of approximately 9% in both SSv2-Full and SSv2-Small when compared to the other methods. Additionally, despite our model having approximately five times the number of parameters as the other methods, the computational load is only about two times higher. This observation suggests that the computational requirements of our model do not increase linearly with the number of parameters, as demonstrated in the last column of Tab. 6. These findings highlight that the computational complexity of our model is not as severe as it might initially appear based solely on the parameter count.

Analysis of Semantic Consistency. In this section, we qualitatively analyze the semantic consistency in the features extracted by our SAFSAR. We compare the t-SNE [44] visualizations of the features extracted with or without the MM Fusion module for five randomly selected actions in HMDB51 and UCF101. As illustrated in Fig. 2, we observe that after incorporating the MM Fusion module, the boundaries among different classes become more distinct and expanded. Simultaneously, the features within the same class display increased compactness. This observation signifies our SAFSAR effectively introduces semantic consistency into the features.



(a) HMDB51, without MM Fusion

(b) HMDB51, with MM Fusion

(c) UCF101, without MM Fusion

(d) UCF101, with MM Fusion

Figure 2. t-SNE plots of the features extracted without or with MM Fusion module in the HMDB51 and UCF101 testing set.

Methods	Params (M)	TFLOPs	TFLOPs / Params
TRX [34]	**47.1**	0.340	0.007
HyRSM [51]	65.6	**0.333**	0.005
MASTAF (R3D) [56]	48.7	0.400	0.008
SAFSAR (ours)	245.4	0.805	**0.003**

Table 6. Computational efficiency analysis for 5-way 1-shot.

5. Conclusion

In this work, we proposed a novel Semantic-aware Few-shot Action Recognition model (SAFSAR) for few-shot action recognition. It was designed to be simple yet effective by leveraging a 3D feature extractor, VideoMAE, and employing an efficient transformer-based multi-modality fusion module for adaptively integrating textual semantics into the video representations. We conducted extensive experiments to evaluate the performance on five few-shot action recognition benchmarks, including SSv2-Full, SSv2-Small, UCF101, HMDB51, and EPIC-KITCHENS. The experimental results demonstrated the efficacy of our proposed SAFSAR model by improving upon existing methods and achieving state-of-the-art performance.

6. Acknowledgements

The authors thank Kaleab Kinfu and Carolina Pacheco for their valuable feedbacks. This work was supported by a Discovery Challenge Award from Johns Hopkins Discovery Fund Program.

References

[1] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. *Advances in neural information processing systems*, 29, 2016. 2

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

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 1, 3

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 1, 3

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

[6] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10618–10627, 2020. 1, 2, 4, 5, 6

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1, 6

[8] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023. 6

[9] Shizhe Chen and Dong Huang. Elaborative rehearsal for zero-shot action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13638–13647, 2021. 4, 6

[10] Zitian Chen, Yanwei Fu, Yinda Zhang, Yu-Gang Jiang, Xiangyang Xue, and Leonid Sigal. Multi-level semantic feature augmentation for one-shot learning. *IEEE Transactions on Image Processing*, 28(9):4594–4605, 2019. 2

[11] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. 6

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4, 6

[13] Guneet S Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. *arXiv preprint arXiv:1909.02729*, 2019. 2

[14] Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: spatially-aware few-shot transfer. *Advances in Neural Information Processing Systems*, 33:21981–21993, 2020. 1, 2

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

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

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

[18] Shreyank N Gowda, Marcus Rohrbach, Frank Keller, and Laura Sevilla-Lara. Learn2augment: learning to composite videos for data augmentation in action recognition. In *European conference on computer vision*, pages 242–259. Springer, 2022. 5

[19] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 6

[20] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3154–3160, 2017. 1

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016. 2

[22] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. *Advances in Neural Information Processing Systems*, 32, 2019. 5

[23] Yifei Huang, Lijin Yang, and Yoichi Sato. Compound prototype matching for few-shot action recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 351–368. Springer, 2022. 3, 5

[24] Rongrong Jin, Xiao Wang, Guangge Wang, Yang Lu, Hai-Miao Hu, and Hanzi Wang. Embedding adaptation network with transformer for few-shot action recognition. In *Asian Conference on Machine Learning*, pages 515–530. PMLR, 2023. 2, 3, 4, 5

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

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

[27] Hongyang Li, David Eigen, Samuel Dodge, Matthew Zeiler, and Xiaogang Wang. Finding task-relevant features for few-shot learning by category traversal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1–10, 2019. 5

[28] Shuyuan Li, Huabin Liu, Rui Qian, Yuxi Li, John See, Mengjuan Fei, Xiaoyuan Yu, and Weiyao Lin. Ta2n: Two-stage action alignment network for few-shot action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1404–1411, 2022. 5

[29] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 3

[30] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2, 4

[31] Khoi D Nguyen, Quoc-Huy Tran, Khoi Nguyen, Binh-Son Hua, and Rang Nguyen. Inductive and transductive few-shot video classification via appearance and temporal alignments. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX*, pages 471–487. Springer, 2022. 5

[32] Xinzhe Ni, Hao Wen, Yong Liu, Yatai Ji, and Yujiu Yang. Multimodal prototype-enhanced network for few-shot action recognition. *arXiv preprint arXiv:2212.04873*, 2022. 2, 3, 4, 5

[33] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2(3):4, 2018. 2

[34] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-relational crosstransformers for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 475–484, 2021. 1, 2, 4, 5, 8

[35] Alexander J Ratner, Henry Ehrenberg, Zeshan Hussain, Jared Dunnmon, and Christopher Ré. Learning to compose domain-specific transformations for data augmentation. *Advances in neural information processing systems*, 30, 2017. 2

[36] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International conference on learning representations*, 2017. 2

[37] Yuheng Shi, Xinxiao Wu, and Hanxi Lin. Knowledge prompting for few-shot action recognition. *arXiv preprint arXiv:2211.12030*, 2022. 2, 3, 4, 5

[38] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 1, 2, 4

[39] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11), 2012. 6

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

[41] Anirudh Thatipelli, Sanath Narayan, Salman Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Bernard Ghanem. Spatio-temporal relation modeling for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19958–19967, 2022. 2, 3, 5

[42] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022. 1, 3, 4, 6

[43] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 3

[44] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 4, 5, 7

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

[47] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *arXiv preprint arXiv:2003.06957*, 2020. 2

[48] Xiao Wang, Weirong Ye, Zhongang Qi, Guangge Wang, Jianping Wu, Ying Shan, Xiaohu Qie, and Hanzi Wang. Task-aware dual-representation network for few-shot action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 3, 4, 5

[49] Xiao Wang, Weirong Ye, Zhongang Qi, Xun Zhao, Guangge Wang, Ying Shan, and Hanzi Wang. Semantic-guided relation propagation network for few-shot action recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 816–825, 2021. 2, 3, 4, 5

[50] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Changxin Gao, Yingya Zhang, Deli Zhao, and Nong Sang. Molo: Motion-augmented long-short contrastive learning for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18011–18021, 2023. 3

[51] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Mingqian Tang, Zhengrong Zuo, Changxin Gao, Rong Jin, and Nong Sang. Hybrid relation guided set matching for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19948–19957, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[52] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 6

[53] Yuyang Wanyan, Xiaoshan Yang, Chaofan Chen, and Changsheng Xu. Active exploration of multimodal complementarity for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6492–6502, 2023. 3

[54] Jiamin Wu, Tianzhu Zhang, Zhe Zhang, Feng Wu, and Yongdong Zhang. Motion-modulated temporal fragment alignment network for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9151–9160, 2022. 3, 5

[55] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12203–12213, 2020. 2

[56] Huanle Zhang, Hamed Pirsiavash, and Xin Liu. Mastaf: A model-agnostic spatio-temporal attention fusion network for few-shot video classification. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2507–2516. IEEE, 2023. 1, 3, 5, 8

[57] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip HS Torr, and Piotr Koniusz. Few-shot action recognition with permutation-invariant attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 525–542. Springer, 2020. 1, 3, 5, 6

[58] Lingling Zhang, Xiaojun Chang, Jun Liu, Minnan Luo, Mahesh Prakash, and Alexander G Hauptmann. Few-shot activity recognition with cross-modal memory network. *Pattern Recognition*, 108:107348, 2020. 2, 3, 5

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

[60] Sipeng Zheng, Shizhe Chen, and Qin Jin. Few-shot action recognition with hierarchical matching and contrastive learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 297–313. Springer, 2022. 5

[61] Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 751–766, 2018. 1, 2, 6

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