Supervised Contrastive Learning for Few-Shot Action Classification

Hongfeng Han[1,3][0000−0003−2385−4335], Nanyi Fei[1,3][0000−0002−3852−9298], Zhiwu Lu[2,3](✉)[0000−0001−6429−7956], and Ji-Rong Wen[2,3][0000−0002−9777−9676]

[1] School of Information, Renmin University of China, Beijing, China
[2] Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[3] Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China {hanhongfeng,feinanyi,luzhiwu,jrwen}@ruc.edu.cn

Abstract. In a typical few-shot action classification scenario, a learner needs to recognize unseen video classes with only few labeled videos. It is critical to learn effective representations of video samples and distinguish their difference when they are sampled from different action classes. In this work, we propose a novel supervised contrastive learning framework for few-shot video action classification based on spatial-temporal augmentations over video samples. Specifically, for each meta-training episode, we first obtain multiple spatial-temporal augmentations for each video sample, and then define the contrastive loss over the augmented support samples by extracting positive and negative sample pairs according to their class labels. This supervised contrastive loss is further combined with the few-shot classification loss defined over a similarity score regression network for end-to-end episodic meta-training. Due to its high flexibility, the proposed framework can deploy the latest contrastive learning approaches for few-shot video action classification. The extensive experiments on several action classification benchmarks show that the proposed supervised contrastive learning framework achieves state-of-the-art performance.

Keywords: Few-shot learning · Contrastive learning · Action classification.

1 Introduction

Recently, the metric-based meta-learning paradigm has led to great advances in few-shot learning (FSL) and become the mainstream [10, 36, 7]. Following such a paradigm, FSL models are typically trained via two learning stages [21]: (1) They are first trained on base classes to learn visual representations, acquiring transferable visual analysis abilities. (2) During the second stage, the models learn to classify novel classes that are unseen before by using only a few labelled samples per novel class. Similar to FSL, contrastive learning (CL) [21] is also deployed to address the labelled data-hungry problem. Specifically, CL is defined as unsupervised or self-supervised learning. The target of CL is to obtain

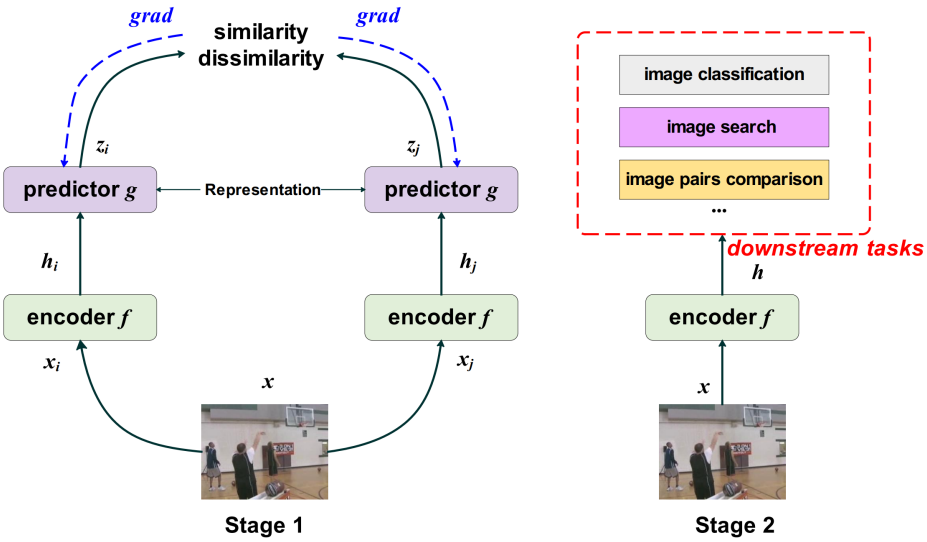


Fig. 1: A typical contrastive learning framework for unsupervised image representation learning. Specifically, two image views x_i and x_j are generated from the same family of image augmentations ($x_i, x_j \sim \Sigma$). A CNN-based encoder network f along with the projection head g is applied to represent each sample effectively. After the network parameters are trained based on a contrastive loss, the projection head g is put away, and only the encoder network f and representations h_i/h_j are used for downstream tasks.

better visual representations to transfer the learned knowledge to downstream tasks such as image classification [29, 8]. As illustrated in Figure 1, a classic CL framework [25, 12, 14, 13] also follows the two learning stages (similar to metric-based meta-learning): (1) An encoder named f and a predictor named g are first trained with constructed positive and negative sample pairs; (2) The learned latent embeddings h_i/h_j are further adapted to downstream tasks of interest. Therefore, it is natural and indispensable to combine FSL and CL.

However, for few-shot action classification, the integration of CL and FSL is extremely challenging because of the complicated video encoding methods. Specifically, two typical methods are widely used: (1) Extracting frame features and then aggregating them. For example, combined with long-short term memory (LSTM), 2D Convolutional Neural Networks (CNNs) are often used for video encoding [32, 20, 40, 5]); (2) Directly extracting spatial-temporal features using 3D CNNs [38, 30, 39, 9, 18] or their variants. For both video encoding practices, the high-level semantic contexts among video frames are difficult to be aligned either spatially or temporally [6, 4].

In this work, we thus propose a novel supervised contrastive learning framework to make a closer integration of CL and FSL for few-shot action classification. Specifically, we first obtain multiple spatial-temporal augmentations

from each video sample for each meta-training episode. Further, we define a supervised contrastive loss over the augmented support samples by constructing positive and negative pairs based on their class labels. Finally, the contrastive loss is combined with a few-shot classification loss defined over a similarity score regression network for the end-to-end episodic meta-training. In addition, the proposed framework can deploy the latest CL methods for few-shot action classification with high flexibility.

In summary, the major contributions of this paper are three-fold:

(1) We devise a spatial-temporal augmentation method to generate different augmentations, facilitating CL to learn better video representations.
(2) We propose a novel supervised contrastive learning framework for few-shot action classification. A similarity score network is shared by both CL and FSL, resulting in a closer integration of the two paradigms.
(3) Extensive experiments on three benchmarks (i.e., HMDB51 [27], UCF101 [34], and Something-Something-V2 [22]) show that the proposed supervised contrastive learning framework achieves state-of-the-art performance.

2 Related Work

Few-shot learning for action classification. Few-shot learning (FSL) approaches are often divided into two main categories: (1) The goal of gradient-based approaches [1, 19, 28, 31] is to achieve rapid learning on a new task with a limited number of gradient update steps while simultaneously avoiding overfitting (which can happen when few labelled samples are used). (2) Metric-based approaches [6, 4, 2, 42, 21] first extract image/video features and then measure the distances/similarities between an embedded query sample and embedded support samples. It is essential to measure the distances in the latent space to determine the class label of query samples. We examine the simplicity and adaptability of the metric-based meta-learning framework in this paper. But note that our proposed video augmentation methods and supervised contrastive learning strategy are also compatible with other few-shot classification solutions.

Contrastive learning. Contrastive learning (CL) is now a relatively new paradigm for unsupervised or self-supervised learning for visual representations, and it has shown some promising results [25, 14, 12, 13, 23, 29, 26, 15, 8]. It is customary for CL methods to learn representations by optimizing the degree to which multiple augmented views of the same data sample agree with one another. This is accomplished by suffering a contrastive loss in the latent embedding space. For example, SimCLR [12] achieves the highest level of agreement possible between various augmented views of the same data sample by obtaining representations and employing a contrastive loss while operating in the latent space. It comes with an improved version called SimCLR v2 [13] that explores larger-sized ResNet models, boosts the performance of the non-linear network (multiple-layer perception, MLP), and incorporates the memory mechanism. Momentum Contrast (MoCo) [25] approach creates a dynamic dictionary using a

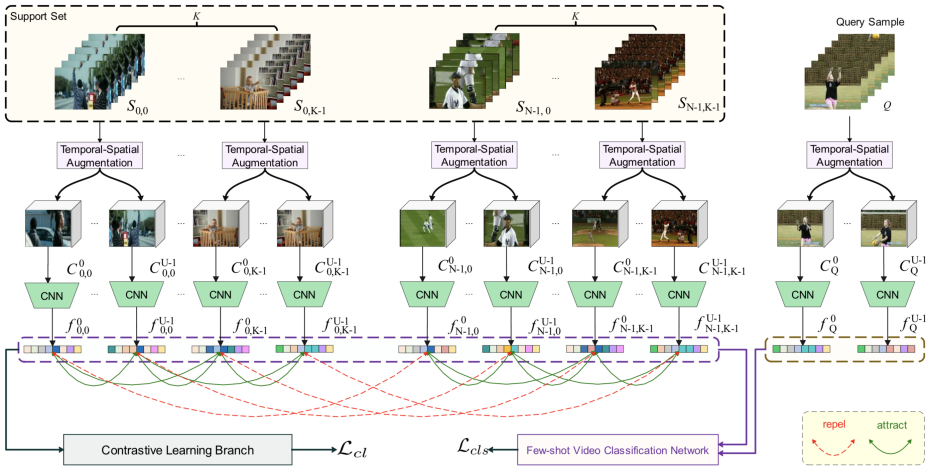


Fig. 2: Architecture of our proposed few-shot action classification framework boosted by supervised contrastive learning. A set of effective spatial-temporal augmentation methods are utilized to generate various video clips (views), which are subsequently fed into the feature extractor (3D CNN) to obtain semantic representation vectors. All these sampled video semantic vectors $f^r_{i,j}$ $(i \in \{0, 1, \cdots, N-1\}, j \in \{0, 1, \cdots, K-1\}, r \in \{0, 1, \cdots, U-1\})$ from the support set are exploited to train a similarity measurement network $\mathcal{M}$ in a supervised way with the contrastive learning loss $\mathcal{L}_{cl}$. Furthermore, $f^r_{i,j}$ together with the representation vectors $f^r_Q, (r \in \{0, 1, \cdots, U-1\})$ of the augmented views of query samples are used to train downstream few-shot classification tasks with softmax loss $\mathcal{L}_{cls}$.

queue structure and a moving-averaged encoder. It allows for the construction of an extensive and consistent dictionary on-the-fly, which enables unsupervised contrastive learning to take place more easily. In this second version [14], the authors apply an MLP-based projection head and more kinds of data augmentation methods to establish strong representations. By performing a stop-gradient operation on one of the two encoder branches, SimSiam [15] is able to optimize the degree to which two augmentations of the same image are similar to one another, which allows it to obtain more meaningful representations even when none of the relevant factors (negative sample pairs, larger batch sizes, or momentum encoders) are present. In this paper, we also evaluate our proposed few-shot video action classification framework with the latest/mainstream CL methods, verifying the flexibility and the independence of our method.

3 Methodology

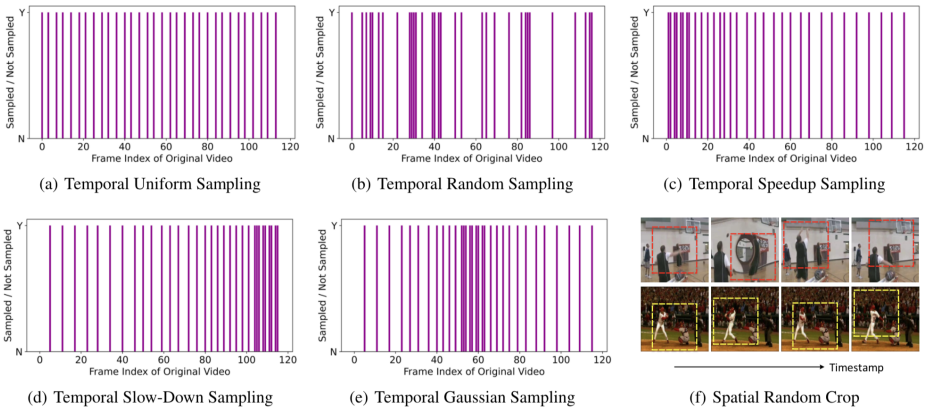
3.1 Framework Overview

To increase the effectiveness of representation ability of the video encoder and measure the similarity score more effectively via contrastive learning, we propose a unified framework that integrates contrastive learning and few-shot learning together in Fig. 2. For an N-way K-shot few-shot episode, video augmentations considering both spatial and temporal dimensions are performed for each video. Concretely, for the j-th input video ($j \in \{0, 1, \cdots, K-1\}$) in the i-th class ($i \in \{0, 1, \cdots, N-1\}$) in the support set (i.e., $S_{i,j}$), we obtain U augmented views/video clips $C^r_{i,j}$ ($r \in \{0, 1, \cdots, U-1\}$). Subsequently, these views with diversity are then followed by a CNN-based feature extractor so that the latent representations can be learned, and outputting the embedded vectors $f^r_{i,j}$. Similarly, for each query sample, we can also obtain the representations of its different augmented views, denoted as f^r_Q ($r \in \{0, 1, \cdots, U-1\}$). Since we have label information for the support set, on the basis of the class labels, we are able to generate positive and negative sample pairs for the purpose of engaging in contrastive learning. That is, two latent vectors belonging to the same class are considered as a positive pair, while they are negative to each other if they come from different classes. In the N-way K-shot scenario with U augmentations, we can generate $N \times U \times (U-1)$ positive pairs and $U^2 \times N(N-1)/2$ negative ones in total (as is illustrated in the dash-lined frame in Fig. 2). Then two branches are extended with the latent vectors: the contrastive learning branch and the few-shot classification branch. The positive and negative pairs are used to train the feature extractor with supervised learning as the input for the contrastive learning branch.

With the loss function defined as $\mathcal{L}_{cl}$, contrastive learning aims to facilitate the feature extractor to generate more discriminative representations, which make positive samples close and negative ones far away in the high-dimensional latent space. As for a few-shot classification scenario, we make use of the mean representation of the K shots for each class (denoted as a prototype) as the class center for the nearest-neighbor search. And a similarity measurement neural network $\mathcal{M}$ is intended to regress the distances between both query samples and prototypes, with the classification softmax loss defined as $\mathcal{L}_{cls}$.

3.2 Supervised Contrastive Learning

For each of the U data augmentation methods, we adopt a combination of temporal and spatial augmentations. The spatial one is the same across all U augmentations, i.e., we perform a random crop in each selected frame (as is shown in Fig. 3(f)). As for the temporal augmentations, we use $U = 5$ methods to provide a diversity of visual representations: uniform sampling, random sampling, speedup sampling, slow-down sampling, and Gaussian sampling. The augmented video clips (views) are further exploited to generate positive and negative sample pairs related to contrastive learning.



(a) Temporal Uniform Sampling (b) Temporal Random Sampling (c) Temporal Speedup Sampling

(d) Temporal Slow-Down Sampling (e) Temporal Gaussian Sampling (f) Spatial Random Crop

Fig. 3: Demonstration of temporal-spatial view augmentations for an input original video with $D = 117$ frames and sampling $T = 32$ frames: (a-e) temporal sampling using uniform, random, speed-up, slow-down, and Gaussian methods, respectively; (f) random spatial crop for each selected frame.

(1) For uniform sampling, let $\mathcal{I}(\sigma)$ denote the frame index of the selected σ-th frame ($\sigma \in \{0, 1, \cdots, T-1\}$, and T is the quantity of selected frames) from the original input video, which follows the distribution defined as:

$$\mathcal{I}(\sigma) \sim \mathcal{U}(0, D), \tag{1}$$

where D represents the total number of the original input video sample, and $\mathcal{U}$ is the uniform distribution.

(2) For random sampling, we directly obtain T frames by independently sampling T times from the original video without any replacement or sorting.

(3) As for speed-up or slow-down sampling, we are motivated by the observation that sometimes meaningful behaviors happen at the front/end along the time dimension in the original video, but which may be ignored by the uniform/random sampling method. The sampled frame $\mathcal{I}(\sigma)$ in both speedup and slow-down cases are defined as:

$$\frac{\mathrm{d}\mathcal{I}(\sigma)}{\mathrm{d}\sigma} = v, \quad \mathcal{I}(0) = 0, \quad \mathcal{I}(T) = D, \tag{2}$$

where v is the sampling velocity which is positive for speedup sampling while negative for the slow-down case. Note that the initial state $\mathcal{I}(0) = 0$ and $\mathcal{I}(T) = D$ limits the range of the sampled index. Speedup sampling samples more frames at the beginning of the input video, and slow-down sampling focus more on frames at the tail.

(4) Gaussian sampling, with slow-down as its first half part and speedup as second half, i.e., it samples most intensively at the middle of a given video sample. Its sampling formulation is the same with Equation (2) but the border state should be initialized as $\mathcal{I}(0) = 0$, $\mathcal{I}(T/2) = D/2$ for the first half and $\mathcal{I}(T/2) =$

Algorithm 1 Supervised Contrastive Learning (SCL)

Require: N, K, video feature extractor g, a set of view augmentations $\mathcal{T}$, batch size B, sampled pair amount M in each batch, similarity measurement network $\mathcal{M}$
Ensure: contrastive learning loss $\mathcal{L}_{cl}$
  for $b \in \{0, 1, \cdots, B-1\}$ **do**
    for sampled video pairs $\{(\boldsymbol{x}_{b,l}, \boldsymbol{x}'_{b,l})\}_{l=0}^{M-1}$ **do**
      Draw two augmentation functions $t \sim \mathcal{T}$, $t' \sim \mathcal{T}$;
      $C_{b,l},\ C'_{b,l} = t(\boldsymbol{x}_{b,l}),\ t'(\boldsymbol{x}'_{b,l})$; # clip generation
      $f_{b,l},\ f'_{b,l} = g(C_{b,l}),\ g(C'_{b,l})$; # representation
      if $(\boldsymbol{x}_{b,l}, \boldsymbol{x}'_{b,l})$ are sampled from the same class **then**
        $y_{b,l} = 1.0$;
      else
        $y_{b,l} = 0.0$;
      end if
    end for
  end for
  for $b \in \{0, 1, \cdots, B-1\}$, $l \in \{0, 1, \cdots, M-1\}$ **do**
    $d_{b,l} = 1.0 - \mathcal{M}(f_{b,l}, f'_{b,l})$; # pairwise distance
  end for
  Update video clip representation network g and similarity measurement network $\mathcal{M}$ by minimizing $\mathcal{L}_{cl}$.

$D/2$, $\mathcal{I}(T) = D$ for the second half. Fig. 3(a-e) illustrate five examples with the same input video sample ($D = 117$) for the five augmentations, respectively ($T = 32$, $v = 4$).

The supervised contrastive learning (SCL) algorithm is summarized in **Algorithm 1**, where the similarity measurement network $\mathcal{M}$ is also shared in the few-shot classification branch, which is used to reflect the distance within each positive/negative pair (the details of $\mathcal{M}$ are described in Section 3.3). We follow the contrastive loss function $\mathcal{L}_{cl}$ used in [11, 24, 35, 44, 16], which is defined as:

$$\mathcal{L}_{cl} = -\frac{1}{BM} \sum_{b=0}^{B-1} \sum_{l=0}^{M-1} y_{b,l} d_{b,l}^2 + (1 - y_{b,l}) \max(m - d_{b,l}, 0)^2, \tag{3}$$

where M is the total constructed positive and negative pairs with a single mini-batch, and $d_{b,l}$ is the distance between two samples of the l-th pair in the b-th input episode, and $y_{b,l}$ is the corresponding ground truth label ($y_{b,l} = 1$ if the pair consists of two views generated from the same class and $y_{b,l} = 1$ otherwise). Note that m is a margin that defines a radius, and the negative pairs affect the loss only when the distance is within this radius.

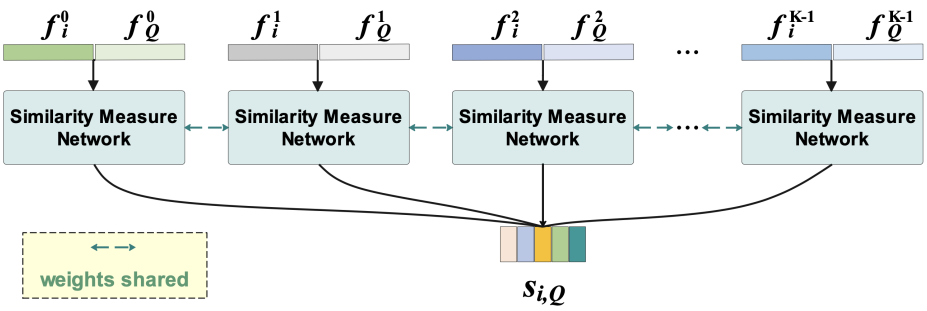


Fig. 4: The schematic illustration of the few-shot action classification process. For the r-th augmented view in the i-th class, the class prototype f_i^r is obtained by averaging the latent representations $f_{i,j}^r$ along the shot dimension j. Together with each query sample's augmented view f_Q^r, the prototype-query pairs are fed into the same similarity measurement network $\mathcal{M}$ which is also used in supervised contrastive learning (see Figure 2) to obtain the final similarity score vector $\mathbf{s}_{i,Q}$.

3.3 Few-Shot Classification

The integration process related to contrastive learning and few-shot learning is reflected in two aspects: (1) The supervised contrastive learning loss is combined with the few-shot classification loss during training. (2) There exists a similarity measurement network $\mathcal{M}$ that is shared across the few-shot classification and the contrastive learning branch to measure the latent distance/similarity between two given augmented views. To exploit the few shots in the support set, we follow Prototypical Network [33] and summarize all shots' latent representations $f_{i,j}^r$ ($i \in \{0, 1, \cdots, N-1\}$, $j \in \{0, 1, \cdots, K-1\}$, $r \in \{0, 1, \cdots, U-1\}$) by computing their average response:

$$f_i^r = \frac{1}{K} \sum_{j=0}^{K-1} f_{i,j}^r. \tag{4}$$

Fig. 4 illustrates the few-shot action classification network. For all the augmented views for a specific class in the support set, the class prototypes f_i^r ($i \in \{0, 1, \cdots, N-1\}$, $r \in \{0, 1, \cdots, U-1\}$) are only concerned with the query sample f_Q^r coming from the same augmentation. The similarity measurement network $\mathcal{M}$ is then utilized to predict the similarity score $s_{i,Q}^r$ between two input views:

$$s_{i,Q}^r = \mathcal{M}(f_i^r, f_Q^r). \tag{5}$$

It is worth mentioning that the similarity score vector $\mathbf{s}_{i,Q}$ of all views is further weighted by a linear layer $\mathbf{w} \in \mathbb{R}^{1 \times U}$, to obtain the final predicted similarity score $s_{i,Q}$ between the i-th class prototype and the query sample:

$$s_{i,Q} = \mathbf{w} \cdot \mathbf{s}_{i,Q}, \tag{6}$$

where $\mathbf{s}_{i,Q} = [s_{i,Q}^0, s_{i,Q}^1, \cdots, s_{i,Q}^{U-1}]^T$.

Finally, a softmax layer maps N similarity scores to a classification distribution vector for each query sample. And the few-shot classification loss $\mathcal{L}_{cls}$ is defined as:

$$\mathcal{L}_{cls} = -\frac{1}{BQ}\sum_{b=0}^{B-1}\sum_{q=0}^{Q-1}\sum_{i=0}^{N-1} y_{b,q,i}\log(\hat{y}_{b,q,i}), \tag{7}$$

where B is the batch size, $y_{b,q,i}$ is the label of the q-th query from the b-th input episode, and $\hat{y}_{b,q,i}$ is the corresponding predicted classification probability.

3.4 Total Learning Objective

We incorporate supervised contrastive learning to the few-shot classification task by adding an auxiliary loss $\mathcal{L}_{cl}$, i.e., the final weighted loss $\mathcal{L}$ is constructed as:

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha\mathcal{L}_{cl}, \tag{8}$$

where α is the balance hyper-parameter.

4 Experiments

4.1 Datasets and Settings

Datasets. In this paper, the proposed supervised contrastive learning framework is evaluated the performance on three different action recognition datasets: HMDB51 [27], UCF101 [34] and Sth-Sth-V2 [22]. **HMDB51** totally contains 6,766 videos distributed in 51 action categories. **UCF101** has included 13,320 videos covering 101 different action-based categories. **Sth-Sth-V2** includes 220,847 videos with 174 different classes. For UCF101 and Sth-Sth-V2, we follow the same splits as in OTAM [6], and they are randomly sampling 64 classes for meta training, 12 classes for meta validation, and 24 classes for meta testing, respectively. For HMDB51, we randomly select 32/6/13 classes for meta training, validation, and testing.

Configuration. It is considered the few-shot scenarios with $N = 5$ and $K = 1, 3, 5$. In each episode, we randomly select N categories, each consisting of K samples as the support set and select another video for each class as the query sample. We train our model over 2,000 episodes and check that the validation set matches an early stopping criterion for every 128 episodes. We use Adam optimizer, and the learning rate is set to 0.001. Furthermore, the average classification accuracies are reported by evaluating 500 and 1000 episodes in the meta-validation and meta-test split, respectively.

3D Backbones. To better demonstrate the generalizability of the proposed framework, we perform extensive experiments with 5 different video feature extraction backbones: C3D [38], R(2+1)D [39], P3D [30], I3D [41] and SlowFast [18]. All backbones are trained with the input size of 224×224. The input clip length for C3D, R(2+1)D, P3D, I3D, and SlowFast are 16, 16, 16,

Table 1: **Comparison to state-of-the-art video action classification approaches** on the HMDB51, UCF101, and Sth-Sth-V2 datasets. All backbones are trained from scratch. Accuracy (%) are reported on average over 1,000 episodes. Note that Neg./Pos. pairs ratio is configured as 2.5.

Methods	Backbone	HMDB51 [27]			UCF101 [34]			Sth-Sth-V2 [22]		
		1-shot	3-shot	5-shot	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot
ARN [43]	C3D	45.53	53.60	59.82	66.60	78.40	84.48	33.44	38.80	45.74
TARN [3]	C3D	66.52	73.30	75.50	85.40	86.72	93.40	38.43	44.54	48.63
ProtoGAN [17]	C3D	35.41	49.89	52.90	61.73	75.89	79.70	33.90	40.72	44.68
FAN [37]	C3D	69.90	71.48	78.20	77.56	87.62	90.80	37.20	43.32	45.82
OTAM [6]	C3D	64.63	79.80	81.90	88.12	91.07	92.10	39.60	47.10	52.30
TAV [4]	C3D	71.30	78.42	83.80	87.90	92.30	92.26	39.40	46.60	49.92
Ours (w/o SCL)	C3D	70.04	77.62	80.51	86.00	90.60	91.20	34.75	41.75	46.28
Ours (full)	C3D	**75.78**	**86.89**	**89.84**	**92.19**	**94.96**	**95.31**	**41.42**	**49.22**	**53.12**

32, and 40 frames, respectively. The global average pooling layer in 3D backbones are remained, and the dimensions of the final clip representation vectors are 4096, 2048, 2048, 2048, and 2304, respectively. All the backbones are trained from scratch. As for the similarity measurement network $\mathcal{M}$, it consists of 5 fully connected layers with 1024, 1024, 512, 512, and 1 neuron.

Contrastive Learning Loss. With contrastive learning enabled, its loss $\mathcal{L}_{cl}$ contributes to the final loss with $\alpha = 1.0$. For the 5-way few-shot action classification scenario, the maximum numbers of generated positive and negative pairs are 100 and 250, respectively. Different positive and negative ratios can be achieved via masking the selected pairs. The margin parameter m in Equation (3) is configured to 0.75 in our work. That is, the distance between two clips of a negative pair is expected to be larger than it.

4.2 Main Results

Comparison to State-of-the-Art. In this paper, we evaluate our proposed architecture with supervised contrastive learning against the action classification methods on *HMDB51* [27], *UCF101* [34] and *Sth-Sth-V2* [22] datasets. Frame-level feature extraction based on 2D CNN and then aggregating them together as the video descriptor is used in original OTAM [6]. For a fair comparison, we change its backbone to C3D to extract feature vectors (each video is split into 16 segments, and each contains 16 frames (clip length)). As for TAV [4], we also re-implement it and replace its 2D backbone with the C3D model, which is then combined with the original temporal structure filter (TSF). For ARN [43], TARN [3], ProtoGAN [17] and FAN [37], we follow the original configurations. The only difference between them and our re-implementation versions is that we train all 3D backbones from scratch rather than use pre-trained weights (such as Kinetics-400) since there inevitably exists a category overlap between mainstream pre-trained models and our evaluation datasets. In Table 1, we summarizes the classification accuracy over 1/3/5 shot(s):

Table 2: **Comparison with different contrastive learning approaches** on the HMDB51, UCF101, and Sth-Sth-V2 datasets. All contrastive learning methods adopt the C3D model (trained from scratch) as their backbones (clip length is 16) to extract video feature vectors. Mean accuracies (%) are reported over 1,000 episodes. Note that Neg./Pos. pair ratio is configured as 2.5.

Methods	HMDB51 [27]			UCF101 [34]			Sth-Sth-V2 [22]		
	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot
FSL	70.04	77.62	80.51	86.00	90.60	91.20	34.75	41.75	46.28
FSL+SCL (MoCo [4])	74.26	78.12	83.22	88.74	91.20	92.51	37.54	44.85	48.96
FSL+SCL (MoCov2 [14])	74.88	85.90	88.60	91.19	93.86	94.30	39.40	48.60	52.04
FSL+SCL (SimCLR [12])	72.32	81.60	84.10	88.90	91.08	92.48	36.90	45.72	49.28
FSL+SCL (SimCLRv2 [13])	74.92	85.20	89.28	91.16	93.66	94.37	40.06	48.90	53.00
FSL+SCL (SimSiam [15])	74.90	85.41	89.17	91.12	93.73	94.70	41.29	48.34	52.92
FSL+SCL (ours)	**75.78**	**86.89**	**89.84**	**92.19**	**94.96**	**95.31**	**41.42**	**49.22**	**53.12**

(1) With supervised contrastive learning disabled, our proposed few-shot classification architecture achieves better performance than ARN [43], Proto-GAN [17] on all three datasets and achieves competitive performance w.r.t. TARN [3] and FAN [37]. However, it performs weaker than OTAM [6] and TAV [4] because both OTAM and TAV mine the temporal alignment information between query and support samples in the latent space, which benefits the subsequent distance measurement and classification.

(2) With supervised contrastive learning enabled, we achieve better classification accuracy in all cases, surpassing prior methods with a significant margin. It illustrates that the auxiliary SCL loss can boost the representation ability and similarity score measurement capacity, resulting in improved final classification accuracy.

(3) Sth-Sth-V2 is much more difficult than HMDB51 and UCF101, as we can observe that the classification results on Sth-Sth-V2 are much lower than those on HMDB51/UCF101 with supervised contrastive learning enabled. Improving classification results on a complex dataset is much more difficult than on simple ones. The difficulty of Sth-Sth-V2 can be further explained by the diversity of samples in each category. For example, the category "putting something onto something" on Sth-Sth-V2 contains many different types of video clips. Almost all labels are general descriptions rather than actions with concrete object names (e.g., not like "putting a cup onto a table"?. The general descriptions increase the classification difficulty significantly.

Contrastive Learning Framework Evaluation. The proposed few-shot action classification architecture with supervised contrastive learning is designed not only for high efficient video representations, but also for pairwise similarity score regression. Therefore, it can adopt other mainstream contrastive learning methods. To demonstrate its generalization ability, MoCo [4], MoCov2 [14], SimCLR [12], SimCLRv2 [13] and SimSiam [15] are compared with our supervised contrastive learning algorithm. The batch size B is 128, and all these models are

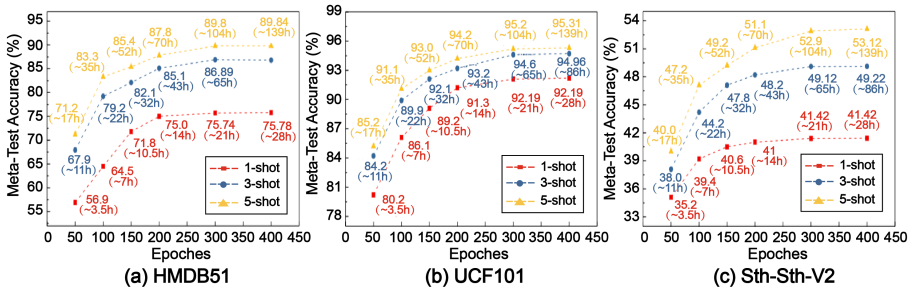


Fig. 5: Model convergence analysis of our proposed supervised contrastive learning algorithm for a few-shot action classification task. Experiments are performed in AWS *ml.g4dn.16xlarge* EC2 instance (64 vCPU and 256G RAM).

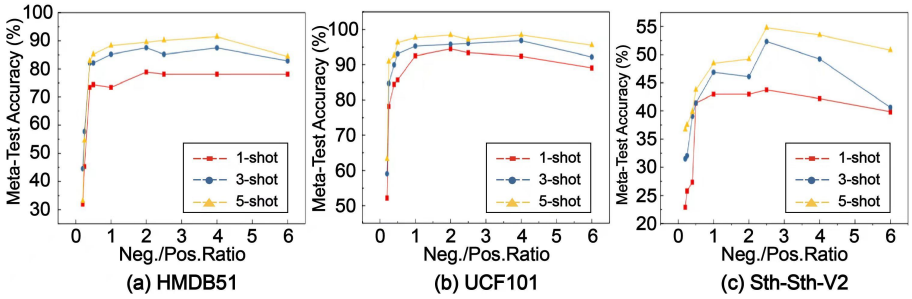


Fig. 6: Comparison of different negative/positive pair ratios for contrastive learning on HMDB51, UCF101, and Sth-Sth-V2 datasets with few-shot action classification. SlowFast is adopted (the speed ratio $\alpha = 8$, and the channel ratio $\beta = 1/8$) as the backbone (clip length is 40).

trained up to 400 epochs. Table 2 shows the few-shot action classification results. As shown in Fig. 5, we show the model convergence curves and training time cost using our SCL algorithm. Experimental results demonstrate that adding the supervised contrastive learning branch indeed improves the few-shot action classification performance. Furthermore, since our proposed SCL algorithm considers an additional similarity network $\mathcal{M}$, it achieves competitive performance boosting.

4.3 Further Evaluations

Different Pos./Neg. Pair Ratios. In the experiment, we evaluate the influence of negative/positive pair ratio in contrastive learning. We configure the ratio to 0.2, 0.25, 0.4, 0.5, 1.0, 2.0, 2.5, 4.0, 6.0 and Fig. 6 plots the average accuracy on 1,000 meta-test episodes using the SlowFast backbone as the video feature extractor. For more details, the speed ratio α is set to 8, and the channel

Table 3: **Comparison of different video representation backbones**. The average classification accuracy (%) with supervised contrastive learning enabled over 1,000 episodes are reported. The values in parentheses represent the percentage improvements over a baseline model with contrastive learning disabled. Note that Neg./Pos. pair ratio is configured to 2.5.

Dataset	K	C3D ([38])	R(2+1)D ([39])	P3D ([30])	I3D ([41])	SlowFast ([18])
HMDB51	1-shot	75.78 (*+5.74*)	76.22 (*+4.78*)	76.84 (*+5.40*)	78.80 (*+3.80*)	78.91 (*+2.10*)
	3-shot	86.89 (*+9.27*)	85.82 (*+6.09*)	86.30 (*+7.60*)	87.52 (*+4.40*)	87.50 (*+3.26*)
	5-shot	89.84 (*+9.33*)	90.02 (*+8.24*)	90.40 (*+6.29*)	91.38 (*+6.10*)	91.41 (*+5.32*)
UCF101	1-shot	92.19 (*+6.19*)	92.60 (*+5.80*)	93.90 (*+6.70*)	94.60 (*+4.65*)	94.53 (*+3.74*)
	3-shot	94.96 (*+4.36*)	95.00 (*+5.96*)	96.38 (*+5.92*)	96.88 (*+5.46*)	96.88 (*+5.28*)
	5-shot	95.31 (*+4.11*)	96.48 (*+3.70*)	97.96 (*+4.26*)	98.50 (*+4.80*)	98.44 (*+3.90*)
Sth-Sth-V2	1-shot	41.42 (*+6.67*)	42.69 (*+6.10*)	43.50 (*+3.29*)	43.74 (*+2.10*)	43.75 (*+2.80*)
	3-shot	49.22 (*+7.47*)	51.00 (*+7.32*)	52.28 (*+3.50*)	52.40 (*+2.46*)	52.34 (*+2.45*)
	5-shot	53.12 (*+6.84*)	53.18 (*+5.43*)	53.93 (*+3.00*)	54.60 (*+2.65*)	54.78 (*+1.58*)

ratio β is 1/8. It is a poor performance of the few-shot action classification when negative/positive pair ratio is smaller than 0.5 on both *HMDB51* and *UCF101* datasets. On the *Sth-Sth-V2* dataset, our model achieves the best results when the ratio is configured to 2.5. From Fig. 6, we can also conclude that unlike SimSiam, our proposed SCL indeed depends on negative samples. One reason is that: not only the video representations are improved (i.e., more discriminative) by contrastive learning, but also the distances between video clips that are essential for few-shot classification are explicitly learned by the contrastive learning loss.

Influence of Different Backbones. To evaluate the generalisability of our proposed framework, we further integrate different video feature extraction backbones. In Table 3, we summarize the few-shot action classification accuracies respectively based on C3D [38], R(2+1)D [39], P3D [30], I3D [41] and SlowFast [18] (the speed ratio $\alpha = 8$, and the channel ratio $\beta = 1/8$) with supervised contrastive learning enabled. The performance improvements are also given in parentheses compared to a simple model with a single few-shot classification branch without contrastive learning. It is clear to see that: (1) For all cases on three different datasets, the proposed framework achieves better results with the supervised contrastive learning branch enabled, which demonstrates the effectiveness as well as the potential for generalization of the methodology that we have developed, i.e., contrastive learning indeed improves the video representation capacity and benefits the distance measurement for classification. (2) The performance improvements are less significant for high-capacity video extraction backbones such as I3D and SlowFast.

Effect of Spatial-Temporal Augmentations. To evaluate the effect of spatial-temporal augmentation methods, we combine different temporal sampling methods with the spatial random crop. In Table 4, we report the performance on HMDB51 with the C3D backbone. We can observe from Table 4 that

Table 4: **Comparison of different combinations of spatial-temporal augmentations** on HMDB51 with C3D as the backbone. Few-shot classification accuracies (%) are reported over 1,000 episodes. Note that Neg./Pos. pair ratio is configured to 2.5.

Augmentation Method	1-shot	3-shot	5-shot
Uniform Samp. (US)	75.00	84.28	87.40
Random Samp. (RS)	74.84	84.10	87.26
Speedup Samp. (SS)	70.42	81.28	83.40
Slow-Down Samp. (SDS)	71.30	82.00	83.36
Gaussian Samp. (GS)	72.60	83.90	86.45
US+RS	74.89	85.27	88.31
US+RS+SS	75.18	85.63	88.99
US+RS+SS+SDS	75.34	86.74	89.70
US+RS+SS+SDS+GS	**75.78**	**86.89**	**89.84**

uniform sampling and random sampling can achieve better performance than speedup, slow-down, or gaussian sampling, which because uniform and random sampling usually obtain the temporal information across the whole time dimension, while for speedup, slow-down, and gaussian sampling, they pay more attention to the beginning, the end and the middle of the video along the time dimension, respectively. Furthermore, combining all these sampling methods together and using learnable weights (attentive) to get the final similarity score (see Fig. 3) will help us mine the video features better.

5 Conclusions

This paper proposes a general few-shot action classification framework powered by supervised contrastive learning, where contrastive learning is deployed to improve the representation quality of videos and a similarity score network is shared by both contrastive learning and few-shot learning to make a closer integration of the two paradigms. Besides, five spatial-temporal video augmentation methods are designed for generating various video sample views in the N-way K-shot few-shot classification scenarios. The significantly improvements achieved by our proposed framework in few-shot action classification is mainly due to: (1) The auxiliary supervised contrastive learning loss makes the video representations more discriminative. (2) The distance measurement between clips is reflected by the similarity score more precisely thanks to a shared similarity score measurement network in both few-shot classification and contrastive learning branches. Importantly, our proposed framework shows strong generalization abilities when different video representation backbones are used. Our proposed framework also has highly flexibility as it can achieve competitive performance when other mainstream contrastive learning approaches are integrated.

Acknowledgements

We would like to thank the anonymous reviewers. This work was supported in part by National Natural Science Foundation of China (61976220 and 61832017), and Beijing Outstanding Young Scientist Program (BJJWZYJH012019100020098).

References

1. Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M.W., Pfau, D., Schaul, T., Shillingford, B., De Freitas, N.: Learning to learn by gradient descent by gradient descent. In: Advances in Neural Information Processing Systems. pp. 3981–3989 (2016)
2. Ben-Ari, R., Nacson, M.S., Azulai, O., Barzelay, U., Rotman, D.: TAEN: Temporal aware embedding network for few-shot action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2786–2794 (2021)
3. Bi., M., Zou., G., Pat., I.: TARN: Temporal attentive relation network for few-shot and zero-shot action recognition. arXiv preprint arXiv:1907.09021 (2019)
4. Bo, Y., Lu, Y., He, W.: Few-shot learning of video action recognition only based on video contents. In: IEEE Winter Conference on Applications of Computer Vision. pp. 584–593 (2020)
5. Brattoli, B., Buchler, U., Wahl, A.S., Schwab, M.E., Ommer, B.: LSTM self-supervision for detailed behavior analysis. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 6466–6475 (2017)
6. Cao, K., Ji, J., Cao, Z., Chang, C.Y., Niebles, J.C.: Few-shot video classification via temporal alignment. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 10615–10624 (2020)
7. Careaga, C., Hutchinson, B., Hodas, N., Phillips, L.: Metric-based few-shot learning for video action recognition. arXiv preprint arXiv:1909.09602 (2019)
8. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Advances in Neural Information Processing Systems. pp. 9912–9924 (2020)
9. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the Kinetics dataset. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4724–4733 (2017)
10. Chen, J., Zhan, L.m., Wu, X.M., Chung, F.l.: Variational metric scaling for metric-based meta-learning. AAAI Conference on Artificial Intelligence pp. 3478–3485 (2020)
11. Chen, J., Yuan, Z., Peng, J., Chen, L., Huang, H., Zhu, J., Liu, Y., Li, H.: DASNet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **14**, 1194–1206 (2020)
12. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. pp. 1597–1607 (2020)
13. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. In: Advances in Neural Information Processing Systems. pp. 22276–22288 (2020)
14. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)

15. Chen, X., He, K.: Exploring simple siamese representation learning. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 15750–15758 (2021)
16. Choi, H., Som, A., Turaga, P.: AMC-Loss: Angular margin contrastive loss for improved explainability in image classification. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 3659–3666 (2020)
17. Dwivedi, S.K., Gupta, V., Mitra, R., Ahmed, S., Jain, A.: ProtoGAN: Towards few shot learning for action recognition. In: IEEE International Conference on Computer Vision Workshop. pp. 1308–1316 (2019)
18. Feichtenhofer, C., Fan, H., Malik, J., He, K.: SlowFast networks for video recognition. In: IEEE International Conference on Computer Vision. pp. 6201–6210 (2019)
19. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning. pp. 1126–1135 (2017)
20. Gammulle, H., Denman, S., Sridharan, S., Fookes, C.: Two stream LSTM: A deep fusion framework for human action recognition. In: IEEE Winter Conference on Applications of Computer Vision. pp. 177–186 (2017)
21. Gidaris, S., Bursuc, A., Komodakis, N., Pérez, P., Cord, M.: Boosting few-shot visual learning with self-supervision. In: IEEE International Conference on Computer Vision. pp. 8059–8068 (2019)
22. 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: IEEE International Conference on Computer Vision. pp. 5842–5850 (2017)
23. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. In: Advances in Neural Information Processing Systems. pp. 21271–21284 (2020)
24. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1735–1742 (2006)
25. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 9726–9735 (2020)
26. Henaff, O.: Data-efficient image recognition with contrastive predictive coding. In: International Conference on Machine Learning. pp. 4182–4192 (2020)
27. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition. In: International Conference on Computer Vision. pp. 2556–2563 (2011)
28. Nichol, A., Achiam, J., Schulman, J.: On first-order meta-learning algorithms. arXiv preprint arXiv:1803.02999 (2018)
29. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
30. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3D residual networks. In: IEEE International Conference on Computer Vision. pp. 5534–5542 (2017)
31. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: International Conference on Learning Representations (2017)
32. Si, C., Chen, W., Wang, W., Wang, L., Tan, T.: An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1227–1236 (2019)

33. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems. pp. 4077–4087 (2017)
34. Soomro, K., Zamir, A.R., Shah, M.: A dataset of 101 human action classes from videos in the wild. Center for Research in Computer Vision **2**(11) (2012)
35. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in Neural Information Processing Systems. pp. 1988–1996 (2014)
36. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H.S., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1199–1208 (2018)
37. Tan, S., Yang, R.: Learning similarity: Feature-aligning network for few-shot action recognition. In: International Joint Conference on Neural Networks. pp. 1–7 (2019)
38. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: IEEE International Conference on Computer Vision. pp. 4489–4497 (2015)
39. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
40. Tsunoda, T., Komori, Y., Matsugu, M., Harada, T.: Football action recognition using hierarchical LSTM. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 155–163 (2017)
41. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 7794–7803 (2018)
42. Wu, Z., Li, Y., Guo, L., Jia, K.: PARN: Position-aware relation networks for few-shot learning. In: IEEE International Conference on Computer Vision. pp. 6658–6666 (2019)
43. Zhang, H., Zhang, L., Qi, X., Li, H., Torr, P., Koniusz, P.: Few-shot action recognition with permutation-invariant attention. In: European Conference on Computer Vision. pp. 525–542 (2020)
44. Zhang, J., Shih, K.J., Elgammal, A., Tao, A., Catanzaro, B.: Graphical contrastive losses for scene graph parsing. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 11527–11535 (2019)