

Few-shot Action Recognition with Permutation-invariant Attention

Hongguang Zhang^{1,2,3,5}, Li Zhang², Xiaojuan Qi^{4,2}, Hongdong Li^{1,5},
Philip H. S. Torr², Piotr Koniusz^{3,1}

¹ Australian National University, Canberra, Australia

² University of Oxford, Oxford, UK

³ Data61/CSIRO, Australia

⁴ The University of Hong Kong, Hong Kong, China

⁵ Australian Centre for Robotic Vision, Australia

Abstract. Many few-shot learning models focus on recognising images. In contrast, we tackle a challenging task of few-shot action recognition from videos. We build on a C3D encoder for spatio-temporal video blocks to capture short-range action patterns. Such encoded blocks are aggregated by permutation-invariant pooling to make our approach robust to varying action lengths and long-range temporal dependencies whose patterns are unlikely to repeat even in clips of the same class. Subsequently, the pooled representations are combined into simple relation descriptors which encode so-called query and support clips. Finally, relation descriptors are fed to the comparator with the goal of similarity learning between query and support clips. Importantly, to re-weight block contributions during pooling, we exploit spatial and temporal attention modules and self-supervision. In naturalistic clips (of the same class) there exists a temporal distribution shift—the locations of discriminative temporal action hotspots vary. Thus, we permute blocks of a clip and align the resulting attention regions with similarly permuted attention regions of non-permuted clip to train the attention mechanism invariant to block (and thus long-term hotspot) permutations. Our method outperforms the state of the art on the HMDB51, UCF101, *mini*MIT datasets.

1 Introduction

Few-shot learning is an open problem with the goal to design algorithms that learn in the low-sample regime. Examples include meta-learning [11, 32, 1, 41, 42], robust feature representations by relation learning [44, 47, 51, 54], gradient-based [59, 39, 43] and hallucination strategies for insufficient data [18, 55].

However, very few papers address video-based few-shot learning. As annotating large video datasets is prohibitive, this makes the problem we study particularly valuable. While results are far from satisfactory on *Kinetics* [3], the largest action recognition dataset, its size of 300,000 video clips with hundreds of frames each exceeds the size of large-scale image ImageNet [38] and Places205 [56] datasets.

There exist few limited works on few-shot learning for action recognition [35, 52, 17, 58]. However, they focus on modeling 3D body joints with graphs [17],

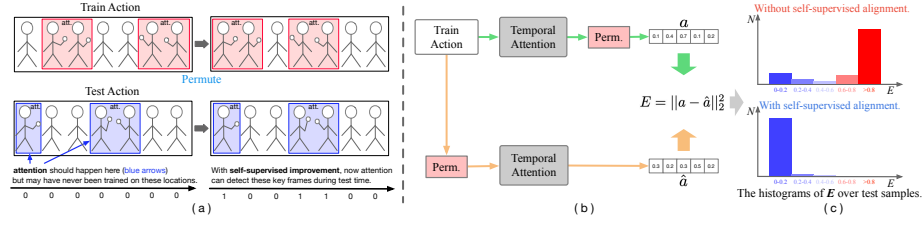


Fig. 1. Augmentation-guided attention by alignment. Fig. 1a shows that discriminative action blocks (in red, top left) may be misaligned with discriminative action blocks of test clip (in blue, bottom left). If the attention unit observes different distributions of locations of discriminative blocks at training and testing time, it fails. With the right approach, one may overcome the distribution shift (top and bottom right panels). Fig. 1b shows how the augmentation-guided attention by alignment works for permutation-based augmentations: (i) we shuffle training blocks of a clip to train an attention on permuted blocks, (ii) we shuffle in the same way coefficients of the attention vector from the non-permuted blocks. Both attention vectors are then encouraged to align by a dedicated loss term during training. Fig. 1c shows histograms of alignment-errors on test data. The top histogram shows larger errors (no alignment loss used in training) while the bottom histogram shows small errors (alignment during testing improves).

attribute-based learning in generative models [35], network design for low-sample classification [52] and salient memory approach [58]. In contrast, we focus on robust relation/similarity, spatial and temporal modeling of short- and long-term range action patterns via permutation-invariant pooling and attention.

To obtain a robust few-shot action recognition approach, we investigate how to: (i) represent discriminative short- and non-repetitive long-term action patterns for relation/similarity learning, (ii) localize temporally discriminative action blocks with limited number of training samples, and (iii) deal with long-term temporal distribution shift of such discriminative patterns (these patterns never re-appear at the same temporal locations even for clips of the same class).

To address the first point, our early experiments indicated that short-term discriminative action patterns can be captured by an encoder with C3D convolutional blocks. Thus, resulting features from a clip undergo permutation-invariant pooling which discards long-term non-repetitive dependencies. Finally, pooled query/support representations form relation descriptors are fed into a comparator.

Regarding the second point, aggregating spatio-temporal blocks with equal weights is suboptimal. Thus, we devise spatial/temporal attention units to emphasize discriminative blocks. Under the low-sample regime, self-supervision by *jigsaw* and *rotation*¹ helps train a more robust encoder, comparator and attention.

¹ Self-supervision assumes generating cheap-to-obtain data (*e.g.*, augmentation by rotations) from the original data and imposing an auxiliary task whose goal is to predict the label of an augmentation with the goal of robust representation learning [5, 4, 16, 20, 19]. We are first to apply the self-supervision by alignment paradigm to the problem of robust attention training (we devise an augmentation-guided attention).

However, vanilla attention (and/or self-supervision) cannot fully promote the invariance to temporal (or spatial) permutations as described next.

To address the third point, we note that long-term dependencies in clips are non-repetitive *e.g.*, videos of the same class often contain relevant action blocks at different temporal locations. Figure 1a shows that discriminative blocks of training and testing clips of *dance* class do not align (top left *vs.* bottom left corner). By permuting the blocks of training (top right), one can make them align with the most discriminative test samples (bottom right). Figure 1b shows that for a given clip, we (i) shuffle its blocks and feed them to the attention mechanism (shuffling pass), (ii) we shuffle accordingly the attention coefficients from a non-shuffled pass through attention, and (iii) we force attention coefficients from both passes to align. Such an attention by alignment deals with the distribution shift of discriminative temporal (and spatial) patterns via *jigsaw* augmentation (but applies also to *rotation*, *zoom*, *etc.*) To summarize, our contributions include:

- i. A robust pipeline with a C3D-based encoder capturing short-term dependencies which yields block representations subsequently aggregated by permutation-invariant pooling into fixed-length representations which form relation descriptors for relational/similarity learning in an episodic setting [47].
- ii. Spatial and temporal attention units which re-weight block contributions during the aggregation step. To improve training of the encoder, comparator and the attention unit under the low-sample regime, we introduce spatial and temporal self-supervision by *rotations*, and *spatial* and *temporal jigsaws*.
- iii. An improved self-supervised attention unit by applying augmentation patterns such as *jigsaws* and/or *rotations* on the input of the attention unit and aligning the output with augmented the same way attention vector coefficients from non-augmented data passed by the attention unit¹. Thus, the attention unit becomes invariant to a given augmentation action by design.
- iv. We propose new data splits for a systematic comparison of few-shot action recognition algorithms and we make them available as existing approaches use each different pipeline concepts, data modality, data splits and protocols².

2 Related work

Below, we discuss zero-, one- and few-shot learning models followed by a discussion on self-supervised learning and second-order pooling.

One- and few-shot learning models have been widely studied in both the shallow [34, 33, 10, 2, 8] and deep learning pipelines [22, 49, 44, 11, 44, 47]. Motivated by the human ability to learn new concepts from few samples, early works [8, 31] employ generative models with an iterative inference. Siamese Network [22] is a two-stream convolutional neural network which generates image descriptors

² Section 2.1 explains that existing works do not specify class/validation splits which yields $\sim 6\%$ variations in accuracy rendering their protocols highly inaccurate. Section 2.2 explains this issue and how we compare our method to existing works.

and learns the similarity between them. Matching Network [49] proposes query-support episodic learning and L -way Z -shot learning protocols³. The similarity between a query and support images is learnt for each episode. At the testing time, each test query (of novel class) is compared against a handful of annotated test support images for rapid recognition. Prototypical Networks [44] compute distances between a datapoint and class-wise prototypes. Model-Agnostic Meta-Learning (MAML) [11] is trained on multiple learning tasks. Relation Net [47] learns relations between query and support images, and it leverages a similarity learning neural network to compare query-support pairs. SalNet [55] uses saliency-guided end-to-end sample hallucination to grow the training set. Graph Neural Networks (GNN) have also been used in few-shot learning [13, 21, 15].

Self-supervised learning leverages free supervision signals residing in images and videos to promote robust representation learning in image recognition [5, 4, 16], video recognition [9, 40, 12], video object segmentation [30, 57] and few-shot image classification [14, 46]. Approaches [16, 4, 5] learn to predict random image rotations, relative pixel positions, and surrogate classes, respectively. Finally, [14, 46] improve few-shot results by predicting image rotations/jigsaw patterns.

Second-order statistics are used by us for permutation-invariant pooling. They are also used for texture recognition [48, 37] by so-called Region Covariance Descriptors (RCD), object and action recognition [26, 50, 25]. Second-order pooling has also been used in fine-grained image classification [28, 7, 50], domain adaptation [24] and the fine-grained few-shot learning [54, 51, 27].

Few-shot action recognition approaches [35, 17, 52] use a generative model, graph matching on 3D coordinates and a dilated networks with class-wise classifiers, respectively. Approach [58] proposes a so-called compound memory network using key-value memory associations. ProtoGAN [6] proposes a GAN model to generate action prototypes to address few-shot action recognition.

2.1 Contrast with existing works

Unlike [17], we use video clips rather than 3D skeletal coordinates. In contrast to [52], we use relation/similarity learning and our training/testing class concepts are disjoint. While [58] memorizes key values/frames, we model short- and long-term dependencies. While [6] forms action prototypes by GAN, we focus on self-supervised attention learning and permutation-invariant aggregation.

In contrast to self-supervision by rotations and jigsaw [14, 46], we use a sophisticated self-supervision on the attention unit for which a dedicated loss performs the alignment between the attention vector of augmented attention unit and the augmented in the same way attention vector from the non-augmented attention unit. Thus, we train a permutation-invariant attention to deal with the distribution shift of discriminative action locations.

Finally, we use second-order pooling [28, 54] for a permutation-invariant aggregation of temporal blocks while [28, 54] work with images. We develop a theory explaining why Power Normalization helps episodic learning.

³ Kindly see [49, 44, 47] for the concept of query, support and episodic learning, and the evaluation protocols which differ from traditional recognition and low-shot learning.

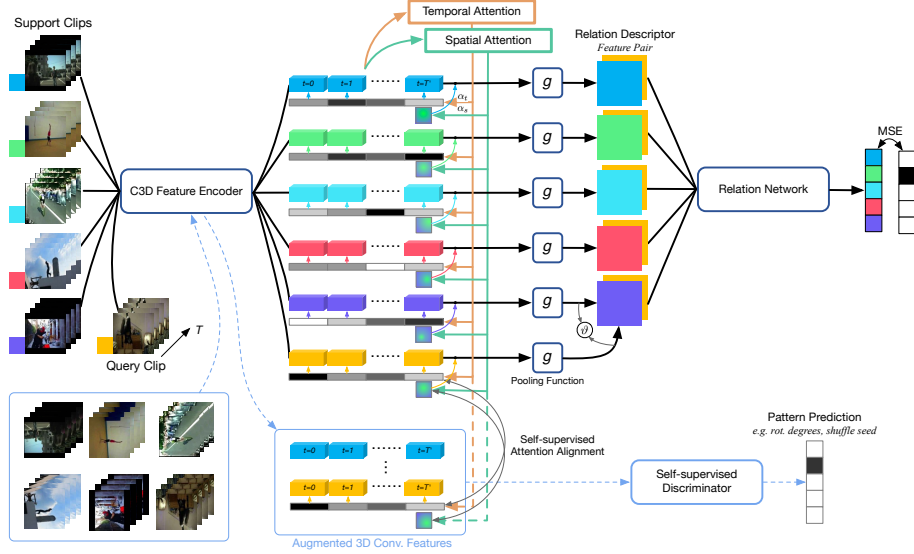


Fig. 2. Our few-shot Action Relation Network (ARN) contains: feature encoder with 4-layer 3D conv. blocks, relation network with 2D conv. blocks, and spatial and temporal attention units which refine the aggregation step. Specifically, we apply second-order pooling (operator g) over encoder outputs (re-weighted by attention vectors) per clip to obtain a Power Normalized Autocorrelation Matrix (AM). Query and support AMs per episode form relation descriptors (by operator ϑ) from which the relation network learns to capture relations. The block (blue dashed line) is the self-supervised learning module which encourages our pipeline to learn auxiliary tasks *e.g.*, *jigsaws*, *rotations*.

2.2 Issues with fair comparisons

Each few-shot action recognition method from Section 2 uses different datasets and evaluation protocols making fair comparisons impossible. Class-wise splits/validation sets are unavailable *ie.*, model [35] uses a random split. Figures 5c and 5d of Section 4 show that the random choice of the split set yields up to $\sim 6\%$ deviation in accuracy rendering such a protocol problematic. Thus, we propose a new protocol with class splits and validation sets made publicly available.

3 Approach

3.1 Pipeline

Figure 2 shows our Action Relation Network (ARN). In contrast to the *Conv-4-64* backbone in few-shot image classification [47, 54, 55], we adopt a C3D-based *Conv-4-64* backbone to extract spatio-temporal features which capture short-range dependencies. Next, we apply second-order pooling on 3D action features re-weighted by attention to obtain second-order statistics which are

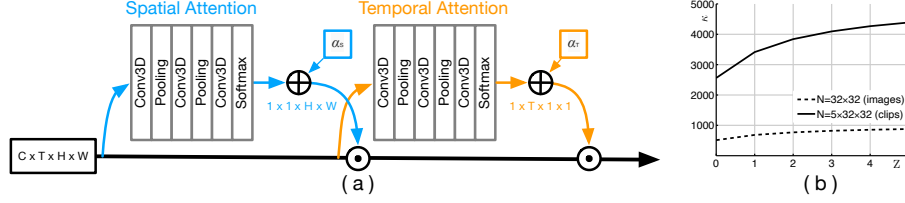


Fig. 3. Spatial and temporal units are shown in Fig. 3a. A naive approach is to directly extract the temporal and spatial attention, whose size is $1 \times T \times H \times W$. However, this is computationally expensive and results in overparametrization. Thus, we split the attention block into separate spatial and temporal branches whose impact is adjusted by α_s and α_t . Fig. 3b is the κ ratio w.r.t. the Z-shot value (see Eq. (6)). The dashed curve shows that as Z grows (0 denotes the regular classification), the memorization burden of co-occurrence (i, j) on the comparator grows κ times for second-order pooling without Power Normalization (as opposed to Power Normalization). The solid line shows that as we use larger N (video clips *vs.* images), not using PN is even more detrimental.

permutation-invariant [28] w.r.t. the spatio-temporal order of features. To paraphrase, we discard the long-range order of temporal (and spatial) blocks captured by the encoder. Finally, second-order matrices form relation descriptors from query/support clips fed into a 2D relation network to capture relations.

Let \mathbf{V} denote a video (*ie.*, with ~ 20 frames) and $\Phi \in \mathbb{R}^{C \times T \times H \times W}$ be features extracted from \mathbf{V} by f :

$$\Phi = f(\mathbf{V}; \mathcal{F}). \quad (1)$$

To aggregate Φ per clip into Ψ , we apply a pooling operator g over the support and query features, resp. For g , we use pooling operators from Sec. 3.2:

$$\Psi = g(\Phi). \quad (2)$$

Once Ψ are computed for query/support clips, they form relation descriptors (via operator ϑ) passed to the relation network r to obtain the relation score ζ_{sq} :

$$\zeta_{sq} = r(\vartheta(\Psi_s, \Psi_q); \mathcal{R}), \quad (3)$$

where \mathcal{R} are parameters of network r , and ϑ forms relation descriptors *e.g.*, we use the concatenation along the channel mode.

We use the Mean Square Error (MSE) loss over support and query pairs:

$$L = \sum_{s \in S} \sum_{q \in Q} (\zeta_{sq} - \delta(l_s - l_q))^2, \quad (4)$$

where $\delta(l_s - l_q) = 1$ if $l_s = l_q$, $\delta(l_s - l_q) = 0$ otherwise. Class labels of support and query action clips are denoted as l_s and l_q .

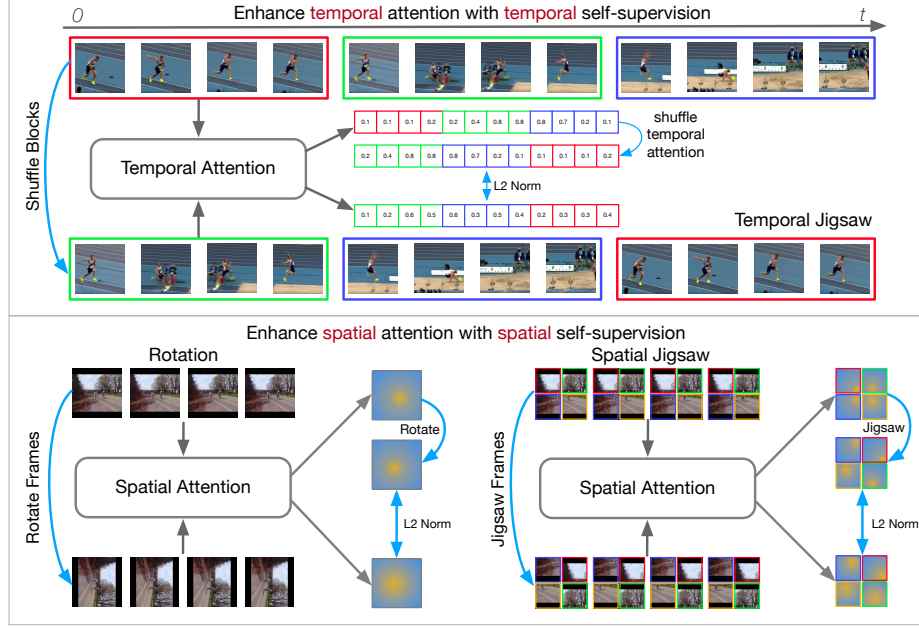


Fig. 4. Augmentation-guided attention by alignment. We firstly collect the encoded representations of original and augmented data, then we extract the temporal or spatial attention vectors from them. We apply the same augmentation(s) on the temporal or spatial attention vectors of the original data resulting in the augmented attention vectors which we align with attention vectors of the augmented data.

3.2 Pooling of encoded representations

For permutation-invariant pooling of temporal (and spatial) blocks, we investigate three pooling mechanisms discussed below.

Average and max pooling are two widely-used pooling functions which can be used for aggregation of $N = T \times W \times H$ fibers (channel-wise vectors) of feature map Φ defined in Eq. (1). The average pooling is given as $\psi = \frac{1}{N} \sum_{n=1}^N \phi_n$ where $\psi \in \mathbb{R}^C$, and $\phi_n \in \mathbb{R}^C$ are N fibers. Similarly, max pooling is given by $\psi_c = \max_{n=1, \dots, N} \phi_{cn}$, $c=1, \dots, C$, and $\psi = [\psi_1, \dots, \psi_C]^T$. Average and max pooling are commutative w.r.t. the input fibers, thus being permutation-invariant. However, first-order pooling is less informative than second-order [23] discussed next.

Second-order pooling captures correlations (or co-occurrences) between pairs of features in N fibres of feature map Φ , which is reshaped such that $\Phi \in \mathbb{R}^{C \times N}$, $N = T \times H \times W$. Such an operator proved robust in classification [23] and few-shot learning [54, 51, 27]. Specifically, we define:

$$\Psi = \eta \left(\frac{1}{N} \sum_{n=1}^N \phi_n \phi_n^T \right) = \eta \left(\frac{1}{N} \Phi \Phi^T \right) \quad \text{where} \quad \eta(\mathbf{X}) = \frac{1 - \exp(\sigma \mathbf{X})}{1 + \exp(\sigma \mathbf{X})}. \quad (5)$$

Matrix $\Psi \in \mathbb{R}^{C \times C}$ is a Power Normalized autocorrelation matrix capturing correlations of fiber features ϕ_n of feature map Φ from Eq. (1) while η applies Power Normalization (PN): we use the zero-centered element-wise Sigmoid [23, 54] on \mathbf{X} , and σ controls the slope of PN. For a given pair of features i and j in matrix Ψ , that is Ψ_{ij} , the role of PN is to detect the likelihood if at least one co-occurrence of features i and j has been detected [23]. According to Eq. (5), second-order pooling is permutation-invariant w.r.t. the order of input fibers as the summation in Eq. (5) is commutative w.r.t. the order of ϕ_1, \dots, ϕ_N . Thus, second-order pooling factors out the spatial and temporal modes of Φ and aggregates clips with varying numbers of temporal blocks (discards the order of long-range spatial/temporal dependencies) into a fixed length representation $\Psi \in \mathbb{R}^{C \times C}$. Below we explain further why second-order pooling with Power Normalization is well suited for episodic few-shot learning.

Relation descriptors between query/support pooled matrices Ψ_q and Ψ_s are formed by operation $\vartheta(\Psi_q, \Psi_s)$ which, in our case, simply performs concatenation of Ψ_q with Ψ_s along the channel mode by $\text{cat}(\Psi_q, \Psi_s) \in \mathbb{R}^{2 \times C \times C}$, and Ψ_s is obtained by the mean (or maximum) along the channel mode between $\Psi_s^1, \dots, \Psi_s^Z$ belonging to the same episode and class ($Z > 1$ for the few-shot case).

It is known from [28] that the Power Normalization in Eq. (5) performs a co-occurrence detection rather than counting (correlation). For classification problems, assume a probability mass function $p_{X_{ij}}(x) = 1/(N+1)$ if $x=0, \dots, N$, $p_{X_{ij}}(x)=0$ otherwise, that tells the probability that co-occurrence between Φ_{in} and Φ_{jn} happened $x=0, \dots, N$ times (given some clip). Note that classification often depends on detecting a co-occurrence (*e.g.*, is there a flower co-occurring with a pot?) rather than counts (*e.g.*, how many flowers and pots co-occur?). Using second-order pooling without PN requires a classifier to observe $N+1$ training samples of *flower and pot* co-occurring in quantities $0, \dots, N$ to memorise all possible co-occurrence count configurations. For relation learning, our ϑ stacks pairs of samples to compare, thus a comparator now has to deal with a probability mass function of $R_{ij} = X_{ij} + Y_{ij}$ depicting *flowers and pots* whose support($p_{R_{ij}}$) = $2N+1 > \text{support}(p_{X_{ij}}) = N+1$ if random variable $X=Y$ (same class). The same is shown by variances $\text{var}(p_{R_{ij}}) > \text{var}(p_{X_{ij}})$. For Z -shot learning, the growth of variance and support equal $(Z+1)N+1$ indicates that the comparator has to memorize more configurations of co-occurrence (i, j) as Z grows.

However, this situation is alleviated by Power Normalization (operator η) whose probability mass function can be modeled as $p_{X_{ij}^\eta}(x) = 1/2$ if $x = \{0, 1\}$, $p_{X_{ij}^\eta}(x) = 0$ otherwise, as PN detects a co-occurrence (or its lack). For Z -shot learning, support($p_{R_{ij}^\eta}$) = $Z+2 \ll \text{support}(p_{R_{ij}}) = (Z+1)N+1$. The ratio given as

$$\kappa = \frac{\text{support}(p_{R_{ij}})}{\text{support}(p_{R_{ij}^\eta})} = \frac{(Z+1)N+1}{Z+2} \quad (6)$$

shows that the comparator has to memorize many more count configurations of co-occurrence (i, j) for naive pooling compared to PN as Z and/or N grow (N depends on the number of temporal and spatial blocks T , H and W). Figure 3 shows how κ varies w.r.t. Z and N . Our modeling assumptions are simple *e.g.*, the

assumption on mass functions with uniform probabilities, the use of the support of mass functions rather than variances to describe variability of co-occurrence (i, j) . Yet, substituting these modeling choices with more sophisticated ones does not affect theoretical conclusions that: (i) PN benefits few-shot learning ($Z \geq 1$) more than the regular classification ($Z = 0$) in terms of reducing possible count configurations of (i, j) , and (ii) for videos (large N) PN reduces the number of count configurations of (i, j) more rapidly than for images (smaller N). While classifiers and comparators do not learn exhaustively all count configurations of co-occurrence (i, j) as they have some generalization ability, they learn quicker and better if the number of count configurations of (i, j) is limited.

3.3 Temporal and spatial attention

Figure 3a introduces decoupled spatial/temporal attention units consisting of three 3D Convolutional blocks and a Sigmoid output layer. Let t and s denote the temporal and spatial attention modules, and the attention be applied ahead of second-order pooling. We obtain temporal and spatial attention maps $\mathbf{T} \in \mathbb{R}^{1 \times T \times 1 \times 1}$ and $\mathbf{S} \in \mathbb{R}^{1 \times 1 \times H \times W}$, and attentive action features Φ^* by:

$$\mathbf{T} = t(\Phi; \mathcal{T}), \mathbf{S} = s(\Phi; \mathcal{S}), \quad (7)$$

$$\Phi^* = (\alpha_t + \mathbf{T}) \cdot (\alpha_s + \mathbf{S}) \cdot \Phi, \quad (8)$$

where \mathcal{T} and \mathcal{S} are network parameters of temporal/spatial attention units while α_t and α_s control the impact of attention vectors.

Using attention helps spot discriminative temporal/spatial blocks, and suppress uninformative regions. However, the attention should be robust to varying distributions of locations of discriminative blocks in clips as proposed below.

3.4 Temporal and spatial self-supervision

Self-supervised Learning (SsL) helps learn representations without using manually-labeled annotations. We impose self-supervision both on encoders and attention units. For temporal self-supervision, we augment clips by shuffling the order of temporal blocks, which primes our network to become robust to long-term non-repetitive temporal dependencies in clips. Self-supervision also helps overcome the low-sample by encouraging network to learn auxiliary tasks. In contrast, previous works shuffled frames which breaks the highly discriminative short-term temporal dependencies. We use the following self-supervision strategies:

- i. **Temporal jigsaw.** Jigsaw, a popular self-supervisory task breaks the object location bias and teaches the network to recognize shuffling. As in [53], we split clips into non-overlapping fixed-length temporal blocks and shuffle them.
- ii. **Spatial jigsaw.** We split frames into four non-overlapping regions, then randomly permute them.
- iii. **Rotation.** As the most popular self-supervisory task are rotations, we uniformly rotate all frames per clip by a random angle (0° , 90° , 180° , 270°).

Figure 2 (blue frame) shows how we apply and recognize the self-supervision patterns *e.g.*, shuffling and rotation angles. Below, we illustrate self-supervision via rotations. Consider the objective function L_{rot} for self-supervised learning with a self-supervision discriminator d , where \mathcal{D} are parameters of d . Thus:

$$\hat{\Phi}_i = f(\text{rot}(\mathbf{V}_i, \theta); \mathcal{F}), \quad (9)$$

$$\mathbf{p}_{rot_i} = d(\hat{\Phi}_i; \mathcal{D}), \quad (10)$$

$$L_{rot} = - \sum_i \log \left(\frac{\exp(\mathbf{p}_{rot_i}[l_{\theta_i}])}{\sum_s \exp(\mathbf{p}_{rot_i}[l_s])} \right), \quad (11)$$

where \mathbf{V}_i is a randomly sampled clip, $\theta \in \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$ is a randomly selected rotation angle of a frame, $l_{\theta_i} \in \{0, 1, 2, 3\}$ is the rot. label for sample i .

Combining the original loss function L with such a self-supervision term L_{rot} results in a self-supervised few-shot action recognition pipeline. However, this objective does not make the attention to be invariant to augmentations per se.

3.5 Augmentation-guided attention by alignment

Figure 4 presents a strategy in which we extract the attention vector for an augmented clip, then we apply the same augmentation to the attention vector obtained from the original non-augmented clip, and we encourage such a pair of augmentation vectors to align by a dedicated MSE loss. This encourages the attention unit to be invariant w.r.t. a given augmentation type. Fig. 1a explains why the temporal permutation strategy benefits few-shot learning while Figure 4 shows how to apply permutations and rotations. As an example, for a rotation-guided spatial-attention we define the alignment loss L_{att} :

$$\hat{\Phi}_i = f(\text{rot}(\mathbf{V}_i, \theta); \mathcal{F}), \quad (12)$$

$$S_i = s(\Phi_i; \mathcal{S}), \hat{S}_i = s(\hat{\Phi}_i; \mathcal{S}), \quad (13)$$

$$L_{att} = \sum_i || \text{rot}(S_i, \theta) - \hat{S}_i ||_F^2 - \lambda || \text{rot}(S_i, \theta) - \hat{S}_i ||_F^2. \quad (14)$$

where λ controls the strictness of alignment. The final objective then becomes:

$$\arg \min_{\mathcal{F}, \mathcal{D}, \mathcal{T}, \mathcal{S}} L + \beta L_{ss} + \gamma L_{att} \quad (15)$$

where β and γ are the hyper-parameters adjusted by cross-validation, L_{ss} is a chosen type of self-supervision *e.g.*, via rotations as introduced in Eq. (11).

4 Experiments

4.1 Experimental setup

Below, we describe our setup and evaluations in detail. To exclude complicated data pre-processing and frame sampling steps typically used in action recognition, we sample uniformly 20 frames along the temporal mode for each dataset.

Table 1. Comparisons between our ARN model and existing works on HMDB51 and UCF101 splits proposed in [35] and a Kinetics split from [58] (given 5-way acc.)

Model	HMDB51 [35]		UCF101 [35]		Kinetics [6]	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
<i>GenApp</i> [35]	—	52.5 ± 3.10	—	78.6 ± 2.1	-	-
<i>ProtoGAN</i> [6]	34.7 ± 9.20	54.0 ± 3.90	57.8 ± 3.0	80.2 ± 1.3	-	-
<i>CMN</i> [58]	-	-	-	-	60.5	78.9
<i>Ours</i>	44.6 ± 0.9	59.1 ± 0.8	62.1 ± 1.0	84.8 ± 0.8	63.7	82.4

Table 2. Ablations of different modules of ARN given our proposed HMDB51 protocol (given 5-way acc.) We used *spatial-jigsaw* for self-supervision.

Baseline	Spatial Attention	Self-supervision	Alignment	1-shot	5-shot
✓				40.83	55.18
✓	✓			41.27	56.12
✓		✓		44.19	58.50
✓	✓	✓		44.61	59.71
✓	✓		✓	43.11	57.35
✓	✓	✓	✓	45.17	60.56

HMDB51 [29] contains 6849 clips divided into 51 action categories, each with at least 101 clips, 31, 10 and 10 classes selected for training, validation and testing.

Mini Moments in Time (miniMIT) [36] contains 200 classes and 550 videos per class. We select 120, 40 and 40 classes for training, validation and testing.

UCF101 [45], action videos from Youtube, has 13320 video clips and 101 action classes. We randomly select 70 training, 10 validation and 21 testing classes.

Kinetics, used by [58] to select a subset for few-shot learning, consists of 64, 12 and 24 training, validation and testing classes. We use it for comparisons.

Training, validation and testing splits on the first three datasets are detailed in our supplementary material while authors of [58] provide the split on Kinetics. The frames of action clips from all datasets are resized to 128×128 . All models are trained on training splits. Validation set is only used for cross-validation.

4.2 Comparison with previous works

Section 2.2 explains the issues with existing methods, protocols, and the lack of publicly available codes. For a fair comparison, we use firstly the protocol of [35] (HMDB51 and UCF101 datasets) but we chose 5 splits at random according to their protocol to average results over multiple runs: we report an average-case result not the best case or a single run result (in contrast to [35]). We also use the Kinetics split of [58], and compare our approach with [35, 58, 6].

Table 1 shows that our ARN (best variant) outperforms GenApp [35], ProtoGAN [6], CMN [58] by a large margin of 3% to 10% on the three protocols. Our standard errors are low as they result from 5 runs on 5 splits (average case) while a large deviation of ProtoGAN [6] was obtained w.r.t. episodes on a single split.

The weakness of protocol [35]. Evaluation protocols in [35] rely on randomly selecting training/testing classes with 50–50 ratio from all classes to form train-

Table 3. Evaluations on HMDB51 (5-way acc.) Attention: Temporal (TA), Spatial (SA). Self-super.: Temp. (TS), Spat. (SS), Self-Super. & Alignment: Temp. (TSA), Spat. (SSA).

Model	1-shot	5-shot
Baseline		
<i>C3D Prototypical Net</i> [44]	38.05 ± 0.89	53.15 ± 0.90
<i>C3D RelationNet</i> [47]	38.23 ± 0.97	53.17 ± 0.86
<i>C3D SoSN</i> [54]	40.83 ± 0.96	55.18 ± 0.86
Temporal/Spatial Attention only (TA vs. SA)		
<i>ARN+TA</i>	41.97 ± 0.97	57.67 ± 0.88
<i>ARN+SA</i>	41.27 ± 0.98	56.12 ± 0.89
<i>ARN+SA+TA</i>	42.41 ± 0.99	56.81 ± 0.87
Temporal/Spatial Self-supervision only (TS vs. SS)		
<i>ARN+TS (temp. jigsaw)</i>	43.79 ± 0.96	58.13 ± 0.88
<i>ARN+SS (spat. jigsaw)</i>	44.19 ± 0.96	58.50 ± 0.86
<i>ARN+SS (rotation)</i>	43.90 ± 0.92	57.20 ± 0.90
Temp./Spat. Self-super. & Att. by alignment (TSA vs. SSA)		
<i>ARN+TSA (temp. jigsaw)</i>	45.20 ± 0.98	59.11 ± 0.86
<i>ARN+SSA (spat. jigsaw)</i>	45.15 ± 0.96	60.56 ± 0.86
<i>ARN+SSA (rotation)</i>	45.52 ± 0.96	58.96 ± 0.87

ing/testing splits on HMDB51 and UCF101. The performance of that protocol varies heavily due to the randomness. Moreover, results of [35] are reported on a single run. Figures 5c and 5d show up to 6% variations due to the randomness, making a fair comparison between models difficult on such a protocol. The lack of validation set makes cross-validation also a random process affecting results.

We rectify all this by providing standardized training, validation, and testing splits on HMDB51, *mini*MIT and UCF101. In what follows, we use our new splits with our few-shot ARN. We equip the Prototypical Net [44], Relation Net [47] and SoSN [54] with a 3D conv. feature encoder (*C3D*) for baselines used below.

ARN modules (ablations). We start by studying ARN modules on HMDB51. Table 2 shows that combining attention with the baseline *C3D SoSN* pipeline brings $\sim 1\%$ gain. Switching to the attention by alignment brings 1.2–1.9% gain over the naive attention unit. Combining self-supervision with (i) the baseline and (ii) the baseline with attention brings $\sim 3\%$ and $\sim 3.5\%$ gain, resp. Combining all units together (attention, self-supervision and alignment) yields $\sim 5\%$ gain. The computational cost is similar to running either self-supervision or alignment.

Thus, in what follows we will report results for the most distinct four variants: (i) baseline (*C3D SoSN*), (ii) Temporal/Spatial Attention only (*TA & SA*), Temporal/Spatial Self-supervision w/o attention (*TS & SS*), and Temporal/Spatial Self-supervision with attention by alignment (*TSA & SSA*).

Pooling (ablations). Section 3.2 discusses pooling variants from Table 5. Second-order pooling (with PN) outperforms second-order pooling (w/o PN) followed by

Table 4. Evaluations on *miniMIT* and UCF101 datasets (given 5-way acc.) See the legend at the top of Table 3 for the description of abbreviations.

Model	miniMIT		UCF101	
	1-shot	5-shot	1-shot	5-shot
<i>C3D Prototypical Net</i> [44]	33.65 \pm 1.01	45.11 \pm 0.90	57.05 \pm 1.02	78.25 \pm 0.73
<i>C3D RelationNet</i> [47]	35.71 \pm 1.02	47.32 \pm 0.91	58.21 \pm 1.02	78.35 \pm 0.72
<i>C3D SoSN</i> [54]	40.83 \pm 0.99	52.16 \pm 0.95	62.57 \pm 1.03	81.51 \pm 0.74
Temporal/Spatial Attention only (TA vs. SA)				
<i>ARN+TA</i>	41.65 \pm 0.97	56.75 \pm 0.93	63.35 \pm 1.03	80.59 \pm 0.77
<i>ARN+SA</i>	41.27 \pm 0.98	55.69 \pm 0.92	63.73 \pm 1.08	82.19 \pm 0.70
<i>ARN+TA+SA</i>	41.85 \pm 0.99	56.43 \pm 0.87	64.48 \pm 1.06	82.37 \pm 0.72
Temporal/Spatial Self-supervision only (TS vs. SS)				
<i>ARN+TS (temp. jigsaw)</i>	42.45 \pm 0.96	54.67 \pm 0.87	63.79 \pm 1.02	82.14 \pm 0.77
<i>ARN+SS (spat. jigsaw)</i>	42.68 \pm 0.95	54.46 \pm 0.88	63.75 \pm 0.98	80.92 \pm 0.72
<i>ARN+SS (rotation)</i>	42.01 \pm 0.94	56.83 \pm 0.86	63.95 \pm 1.03	81.09 \pm 0.76
Temp./Spat. Self-super. & Att. by alignment (TSA vs. SSA)				
<i>ARN+TSA (temp. jigsaw)</i>	42.65 \pm 0.94	57.35 \pm 0.85	65.46 \pm 1.05	82.97 \pm 0.71
<i>ARN+SSA (spat. jigsaw)</i>	42.92 \pm 0.95	56.21 \pm 0.85	66.04 \pm 1.01	82.68 \pm 0.72
<i>ARN+SSA (rotation)</i>	43.05 \pm 0.97	56.71 \pm 0.87	66.32 \pm 0.99	83.12 \pm 0.70

average and max pooling. Combining average pooling with PN boosts its results which is consistent with the theoretical analysis in Figure 3. In what follows, we use the best pooling variant only, that is second-order pooling with PN.

Main evaluations. Tables 3 and 4 present main evaluations on the proposed by us protocols. Notably, our approaches outperform all baselines (known approaches enhanced by us with the C3D-based encoder). Below, we detail the results.

Attention. Tables 3 and 4 investigate the Temporal and Spatial Attention denoted as (*TA*) and (*SA*) on our few-shot ARN. TA on the 1-shot and 5-shot protocols improves the accuracy by 1.0% and 2.5% while SA boosts the 1- and 5-shot accuracy by 0.5% and 1.0%, respectively. For the combined Temporal and Spatial Attention (*TA+SA*), the Eq. (8) is used with $\alpha_s = 1.0$ and $\alpha_t = 0.5$ (HMDB51) and $\alpha_s = 1.5$ and $\alpha_t = 1.0$ (UCF101) chosen on the validation split. Tables 3 and 4 show that SA+TA achieves a further improvement of up to 1.1% for 1-shot learning but for 5-shot learning it may suffer an 0.8% drop in accuracy compared to the best score of TA and SA while still achieving between an 0.8 and 4.2% gain over the baseline C3D SoSN. This is consistent with our argument that vanilla attention units can be further improved for a better performance.

Table 5. Comparison of pooling functions on our HMDB51 split (5-way 1-shot).

No Pooling	Average	Average+PN	Max	Second-order (w/o PN)	Second-order (with PN)
35.71	39.51	40.02	38.95	39.97	40.83

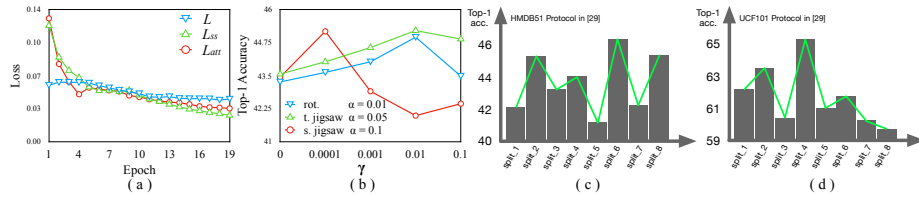


Fig. 5. In Fig. 5a are the loss curves for Eq. (15). Fig. 5b shows the validation score w.r.t. γ (1-shot prot.) Applying Spatial Self-super. & Attention by alignment (SSA) $\gamma > 0$ outperforms the Spatial Self-super. & Attention only ($\gamma = 0$). Fig. 5c and 5d show the performance variation on random splits of HMDB51 and UCF101 proposed by [35].

Temporal/spatial self-supervision. In this experiment, we disable attention units. Tables 3 and 4 show that self-supervision w.r.t. either temporal or spatial mode boosts performance of 1-shot and 5-shot learning over the C3D SoSN baseline on HMDB51 up to 3.2%. On miniMIT, we observe gains between 1.2 and 4.6%. On UCF101, we see gains between 0.6 and 1.4%. However, for UCF101 dataset, self-supervision by the spatial jigsaw and rotation lead to a marginal performance drop on 5-shot learning compared to C3D SoSN.

Temporal/spatial self-supervision & attention by alignment. According to Tables 3 and 4, the gains are in 2–5% range compared to the baseline C3D SoSN. Figure 5a shows the training loss w.r.t. epoch (HMDB51) (*temp. jigsaw*). Figure 5b shows the validation accuracy (HMDB51) w.r.t. γ for SS (*rot.*) As can be seen, Self-supervision combined with Attention by alignment (any curve for $\gamma > 0$) scores higher than Self-supervision with Attention only ($\gamma = 0$).

5 Conclusions

We have proposed a new few-shot Action Recognition Network (ARN) which comprises an encoder, comparator and an attention mechanism to model short- and long-range temporal patterns. We have investigated the role of self-supervision via spatial and temporal augmentations/auxiliary tasks. Moreover, we have proposed a novel mechanism dubbed *attention by alignment* which tackles the so-called distribution shift of temporal positions of discriminative long-range blocks. By combining losses of self-supervision and attention by alignment, we see gains of up to 6% accuracy. We make our dataset splits publicly available to facilitate fair comparisons of few-shot action recognition pipelines.

Acknowledgements. This research is supported in part by the Australian Research Council through Australian Centre for Robotic Vision (CE140100016), Australian Research Council grants (DE140100180), the China Scholarship Council (CSC Student ID 201603170283). Hongdong Li is funded in part by ARC-DP (190102261) and ARC-LE (190100080). We thank CSIRO Scientific Computing, NVIDIA (GPU grant) and the National University of Defense Technology.

References

1. Antoniou, A., Edwards, H., Storkey, A.: How to train your maml. arXiv preprint (2018)
2. Bart, E., Ullman, S.: Cross-generalization: Learning novel classes from a single example by feature replacement. In: CVPR (2005)
3. Carreira, J., Zisserman, A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In: CVPR (2018)
4. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: CVPR (2015)
5. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. In: NeurIPS (2014)
6. Dwivedi, S.K., Gupta, V., Mitra, R., Ahmed, S., Jain, A.: Protogan: Towards few shot learning for action recognition. arXiv preprint (2019)
7. Engin, M., Wang, L., Zhou, L., Liu, X.: Deepkspd: Learning kernel-matrix-based SPD representation for fine-grained image recognition. In: ECCV. vol. 11206, pp. 629–645 (2018). https://doi.org/10.1007/978-3-030-01216-8_38
8. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. TPAMI (2006)
9. Fernando, B., Bilen, H., Gavves, E., Gould, S.: Self-supervised video representation learning with odd-one-out networks. In: CVPR (2017)
10. Fink, M.: Object classification from a single example utilizing class relevance metrics. In: NeurIPS (2005)
11. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML (2017)
12. Gan, C., Gong, B., Liu, K., Su, H., Guibas, L.J.: Geometry guided convolutional neural networks for self-supervised video representation learning. In: CVPR (2018)
13. Garcia, V., Bruna, J.: Few-shot learning with graph neural networks. In: ICLR (2018)
14. Gidaris, S., Bursuc, A., Komodakis, N., Pérez, P., Cord, M.: Boosting few-shot visual learning with self-supervision. In: ICCV (2019)
15. Gidaris, S., Komodakis, N.: Generating classification weights with gnn denoising autoencoders for few-shot learning. In: CVPR (2019)
16. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. arXiv preprint (2018)
17. Guo, M., Chou, E., Huang, D.A., Song, S., Yeung, S., Fei-Fei, L.: Neural graph matching networks for fewshot 3d action recognition. In: ECCV (2018)
18. Hariharan, B., Girshick, R.: Low-shot visual recognition by shrinking and hallucinating features. In: ICCV (2017)
19. Jian, S., Hu, L., Cao, L., Lu, K.: Representation learning with multiple lipschitz-constrained alignments on partially-labeled cross-domain data. In: AAAI. pp. 4320–4327 (2020)
20. Jian, S., Hu, L., Cao, L., Lu, K., Gao, H.: Evolutionarily learning multi-aspect interactions and influences from network structure and node content. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 598–605 (2019)
21. Kim, J., Kim, T., Kim, S., Yoo, C.D.: Edge-labeling graph neural network for few-shot learning. In: CVPR (2019)
22. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML Deep Learning Workshop (2015)

23. Koniusz, P., Cherian, A., Porikli, F.: Tensor representations via kernel linearization for action recognition from 3d skeletons. In: ECCV (2016)
24. Koniusz, P., Tas, Y., Porikli, F.: Domain adaptation by mixture of alignments of second-or higher-order scatter tensors. In: CVPR (2017)
25. Koniusz, P., Wang, L., Cherian, A.: Tensor representations for action recognition. TPAMI (2020)
26. Koniusz, P., Yan, F., Gosselin, P.H., Mikolajczyk, K.: Higher-order occurrence pooling for bags-of-words: Visual concept detection. TPAMI (2017)
27. Koniusz, P., Zhang, H.: Power normalizations in fine-grained image, few-shot image and graph classification. TPAMI (2020)
28. Koniusz, P., Zhang, H., Porikli, F.: A deeper look at power normalizations. In: CVPR (2018)
29. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition. In: ICCV (2011)
30. Lai, Z., Lu, E., Xie, W.: Mast: A memory-augmented self-supervised tracker. In: CVPR (2020)
31. Lake, B.M., Salakhutdinov, R., Gross, J., Tenenbaum, J.B.: One shot learning of simple visual concepts. CogSci (2011)
32. Lee, K., Maji, S., Ravichandran, A., Soatto, S.: Meta-learning with differentiable convex optimization. In: CVPR (2019)
33. Li, F.F., VanRullen, R., Koch, C., Perona, P.: Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences* (2002)
34. Miller, E.G., Matsakis, N.E., Viola, P.A.: Learning from one example through shared densities on transforms. In: CVPR (2000)
35. Mishra, A., Verma, V.K., Reddy, M.S.K., Arulkumar, S., Rai, P., Mittal, A.: A generative approach to zero-shot and few-shot action recognition. In: WACV (2018)
36. Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S.A., Yan, T., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C., Oliva, A.: Moments in time dataset: one million videos for event understanding. TPAMI (2019)
37. Romero, A., Terán, M.Y., Gouiffès, M., Lacassagne, L.: Enhanced local binary covariance matrices (ELBCM) for texture analysis and object tracking. MIRAGE (2013)
38. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. IJCV (2015)
39. Rusu, A.A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., Hadsell, R.: Meta-learning with latent embedding optimization. In: ICLR (2019)
40. Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S.: Time-contrastive networks: Self-supervised learning from pixels. In: ICRA (2017)
41. Simon, C., Koniusz, P., Nock, R., Harandi, M.: Deep subspace networks for few-shot learning. In: NeurIPS workshops (2019)
42. Simon, C., Koniusz, P., Nock, R., Harandi, M.: Adaptive subspaces for few-shot learning. In: CVPR (2020)
43. Simon, C., Koniusz, P., Nock, R., Harandi, M.: On modulating the gradient for meta-learning. In: ECCV (2020)
44. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: NeurIPS (2017)
45. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint (2012)

46. Su, J.C., Maji, S., Hariharan, B.: Boosting supervision with self-supervision for few-shot learning. arXiv preprint (2019)
47. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: CVPR (2018)
48. Tuzel, O., Porikli, F., Meer, P.: Region covariance: A fast descriptor for detection and classification. In: ECCV (2006)
49. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: NeurIPS (2016)
50. Wang, L., Zhang, J., Zhou, L., Tang, C., Li, W.: Beyond covariance: Feature representation with nonlinear kernel matrices. In: ICCV. pp. 4570–4578 (2015). <https://doi.org/10.1109/ICCV.2015.519>
51. Wertheimer, D., Hariharan, B.: Few-shot learning with localization in realistic settings. In: CVPR (2019)
52. Xu, B., Ye, H., Zheng, Y., Wang, H., Luwang, T., Jiang, Y.G.: Dense dilated network for few shot action recognition. In: ICMR (2018)
53. Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., Zhuang, Y.: Self-supervised spatiotemporal learning via video clip order prediction. In: CVPR (2019)
54. Zhang, H., Koniusz, P.: Power normalizing second-order similarity network for few-shot learning. In: WACV (2019)
55. Zhang, H., Zhang, J., Koniusz, P.: Few-shot learning via saliency-guided hallucination of samples. In: CVPR (2019)
56. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. NeurIPS (2014)
57. Zhu, F., Zhang, L., Fu, Y., Guo, G., Xie, W.: Self-supervised video object segmentation. arXiv preprint (2020)
58. Zhu, L., Yang, Y.: Compound memory networks for few-shot video classification. In: ECCV (2018)
59. Zintgraf, L., Shiarli, K., Kurin, V., Hofmann, K., Whiteson, S.: Fast context adaptation via meta-learning. In: ICML (2019)