Unsupervised Few-Shot Action Recognition via Action-Appearance Aligned Meta-Adaptation

Jay Patravali[⋆‡] Gaurav Mittal[⋆†] Ye Yu[†] Fuxin Li[‡] Mei Chen[†]
[†]Microsoft [‡]Oregon State University
{gaurav.mittal, yu.ye, mei.chen}@microsoft.com {patravaj,lif}@oregonstate.edu

Abstract

*We present MetaUVFS as the first **U**nsupervised **Meta**-learning algorithm for **V**ideo **F**ew-**S**hot action recognition. MetaUVFS leverages over 550K unlabeled videos to train a two-stream 2D and 3D CNN architecture via contrastive learning to capture the appearance-specific spatial and action-specific spatio-temporal video features respectively. MetaUVFS comprises a novel Action-Appearance Aligned Meta-adaptation (A3M) module that learns to focus on the action-oriented video features in relation to the appearance features via explicit few-shot episodic meta-learning over unsupervised hard-mined episodes. Our action-appearance alignment and explicit few-shot learner conditions the unsupervised training to mimic the downstream few-shot task, enabling MetaUVFS to significantly outperform all state-of-the-art unsupervised methods on few-shot benchmarks. Moreover, unlike previous few-shot action recognition methods that are supervised, MetaUVFS needs neither base-class labels nor a supervised pretrained backbone. Thus, we need to train MetaUVFS just once to perform competitively or sometimes even outperform state-of-the-art supervised methods on popular HMDB51, UCF101, and Kinetics100 few-shot datasets.*

1. Introduction

Few-shot learning [36, 53, 61, 17, 51, 48, 36, 14, 10, 27, 66] has emerged as a school of approaches that train a model to transfer-learn or adapt quickly on novel, often out-of-domain, classes using as few labeled samples as possible to mitigate the lack of large-scale supervision for these novel classes. Few-shot learning is highly relevant for videos because collecting large-scale labeled video data is extra challenging with the additional temporal dimension. There has been work utilizing both 2D and 3D CNNs [74, 5, 15, 68, 4, 71] to achieve strong results on few-shot action recognition in videos. However, these are supervised approaches and require large amounts of labeled base-class data and/or large-scale supervised pretrained backbones [5, 4, 15, 68] that are not only prohibitively expensive to scale but also oftentimes unattainable. Meanwhile, there is virtually infinite unlabeled video data at our disposal through the rise of multi-media social networking. This motivates us to address the question, *"Can we develop models for video action recognition that perform competitively on few-shot benchmarks without the use of either base-class labels or any external supervision?"*

Existing unsupervised video representation learning methods [47, 55, 24] provide task-agnostic representations that apply to various downstream tasks. However, as we

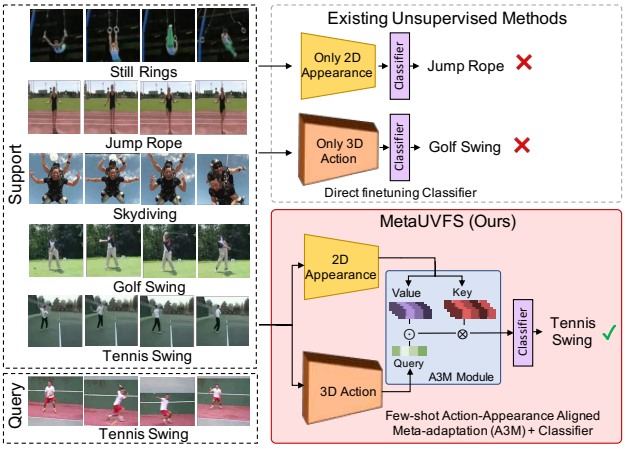

Figure 1. The above example shows a 1-support 5-way few-shot video action recognition task to classify a query sample of novel *Tennis Swing* class. Using only appearance with a 2D CNN incorrectly predicts *Jump Rope* as it relies on only frame-level spatial cues. Using only action with a 3D CNN incorrectly predicts *Golf Swing* as it matches based on just the swinging action without paying attention to the spatial cues. Whereas MetaUVFS predicts the correct class via the Action-Appearance Aligned Meta-adaptation (A3M) module that learns to align and relate the action with the appearance attributes via few-shot meta-learning. All three methods are trained using unlabeled videos.

⋆ Authors with equal contribution.
Work done when Jay Patravali was a research intern at Microsoft.

show in later sections, these methods are not specialized for the few-shot learning task with novel classes and therefore perform sub-optimally on them.

To this end, we propose MetaUVFS as the first method for unsupervised meta-learning for few-shot video action recognition. MetaUVFS leverages large-scale (over half a million) unlabeled video data to learn video representations via contrastive learning and then trains an explicit few-shot meta-learner using episodes that are hard-mined over the learned representations. The episodic meta-learning helps mimic the episodic few-shot meta-testing during the training phase. This imposes a downstream task-specific prior on the learned video representations and reduces the knowledge gap between training and testing.

We introduce an unsupervised two-stream action-appearance network in MetaUVFS to learn fine-grained spatio-temporal 3D features over video segments via an action stream and spatial 2D features over video frames via an appearance stream. Direct finetuning of either feature alone can be sub-optimal in a challenging few-shot scenario as illustrated in Fig. 1. Instead, we design an Action-Appearance Aligned Meta-adaptation module (A3M) in the few-shot meta-learner of MetaUVFS that combines the two streams by learning a spatio-temporal alignment of appearance over action features. A3M learns an attention map conditioned on the action and appearance features to better focus on the action-specific features in the frame-level appearance embeddings. This helps to improve intra-class similarity and reduce inter-class confusion for few-shot.

Consequently, MetaUVFS outperforms all state-of-the-art (SoTA) unsupervised video learning methods on multiple benchmark datasets and also outperforms or performs competitively against the SoTA few-shot action recognition methods. To summarize, our main contributions are,

1. We propose MetaUVFS as the first unsupervised meta-learning algorithm for few-shot video action recognition.

2. MetaUVFS uses a two-stream network to learn action and appearance-specific features via contrastive learning over 550K unlabeled videos. It employs a novel Action-Appearance Aligned Meta-adaptation (A3M) module that is episodically trained via hard-mined episodes to specialize for few-shot downstream tasks.

3. MetaUVFS outperforms all SoTA unsupervised methods across multiple few-shot benchmarks and performs competitively to or even outperforms some of the SoTA few-shot action recognition methods.

2. Related Work

Supervised Few-shot Learning A typical supervised few-shot learning setting has a set of *base*-classes with a large number of labeled samples and a set of *novel* classes with few labeled samples (not enough for plain finetuning). It is evaluated in a *meta-testing* phase where it classifies samples (query) from the novel classes based on a few, e.g. 1 or 5 labeled examples (support).

For images, few-shot learning approaches include metric-learning based [51, 61, 53] that learn to minimize the distance between support and query embeddings or optimization based [14, 48] that develop rapidly learnable models for efficient adaptation on novel classes. Using just the base-class data inhibits generalization to novel classes. There are, therefore, approaches using data augmentation/hallucination [66, 27] or simply training larger supervised models with larger dataset with non-episodic few-shot learning [65, 10, 68]. There are also few-shot approaches using some form of attention/alignment module for improved performance [17, 31, 12] but these are image-specific and are not compatible with the action-appearance features aligned by our A3M module.

To the best of our knowledge, existing few-shot learning work for videos are all supervised approaches. Proto-GAN [38] uses GANs [20] to synthesize addition examples for novel classes, CMN [74] uses memory augmented networks [50] to store video features for query matching, and R-3DFSV [68] uses a large pretrained 3D CNN along with weak labels to augment novel class support samples. There is also work using different forms of cross-attention/alignment such as TARN and ARN [3, 71] capturing spatio-temporal dependencies via attention, OTAM [5] matching query-support pairs via metric-learning based temporal alignment, RVN [4] aligning support-query features via LSTMs, and AMeFu-Net [15] aligning appearance and motion by fusing depth with RGB. Some methods also leverage auxiliary self-supervision to boost few-shot performance [16, 49, 12, 71]. However, unlike previous formulations that either align support and query or use additional modality along with being supervised, our A3M module in MetaUVFS learns to align 2D and 3D features using hard-mined episodes in a purely unsupervised manner.

Supervised Action Recognition Previous methods use either 2D CNNs with frame-level features [19, 13] or 3D CNNs [25, 57] with spatio-temporal features for supervised action recognition. 2D models suffer from the lack of long-term temporal reasoning while 3D models tend to overfit due to larger parameter count. To mitigate this, recent methods introduce self-attention [64], temporal relation [73], factorized 3D convolutions [58], 2D replacements [69], multi-grid scheduler [67] and slow-fast networks [13]. There are also two-stream networks using both 2D and 3D CNNs [63, 7, 54] exploiting optic flow or frame residuals with RGB that we take inspiration from to design our novel action-appearance two-stream network to learn from unlabeled videos.

Unsupervised Few-Shot Learning Recently, unsupervised meta-learning approaches for few-shot image classification [34, 42, 32] have shown competitive performance without using base-class labels or external supervision. Our MetaUVFS drew inspiration from these works to leverage unlabeled videos for few-shot action recognition.

Unsupervised Video Representation Learning Solving pretext tasks in images [11, 72, 18, 44] has inspired methods to learn from unlabeled videos [33, 35] via pretext tasks such as sorting frames and predicting video speed [43, 39, 70, 62, 2, 41, 22, 46]. Recently, methods using contrastive learning (InfoNCE) [45] have been the most effective in harnessing large-scale unlabeled data [28, 8, 47, 55, 24] and perform comparably to supervised methods on vision tasks. Although these methods have shown low-shot learning capabilities, it is primarily limited to being able to finetune on *in-distribution* training classes with a tiny fraction of full-labeled dataset. Unlike the proposed MetaUVFS, without any dedicated few-shot meta-learning mechanism during training, the existing unsupervised methods still require a full-size labeled dataset to optimally transfer to a downstream task with *out-of-distribution* novel classes.

3. Method

We first describe the unsupervised training of the two-stream network in MetaUVFS. We then explain the unsupervised few-shot meta-training and testing of MetaUVFS.

3.1. Two-stream Video Networks

As shown in Fig. 2a, MetaUVFS has a 2D CNN-based *appearance stream* $f^{ap}(\cdot)$ that captures the high-level spatial semantics of the video. $f^{ap}(\cdot)$ encodes a sequence of F frames, $\mathbf{X^{ap}} = [x_t^{ap}]_{t=1}^{F}$, into embeddings $\mathbf{h^{ap}} = [h_t^{ap}]_{t=1}^{F}$ where $h_t^{ap} = f^{ap}(x_t^{ap})$. $\mathbf{h^{ap}}$ are averaged to obtain $\bar{h}^{ap}$. MetaUVFS also has a 3D CNN-based *action stream* $f^{act}(\cdot)$ that captures the spatio-temporal semantics of the video. $f^{act}(\cdot)$ encodes another F' frames, $\mathbf{X^{act}} = [x_t^{act}]_{t=1}^{F'}$, into a single embedding h^{act}, where $h^{act} = f^{act}(\mathbf{X^{act}})$. Inductive biases of using 2D and 3D convolutional kernels in the appearance and action streams respectively enable the streams to specialize in capturing the appearance and action-related video information.

3.2. Two-stream Unsupervised Training Objective

The training objective of our two-streams network is based on the multi-view InfoNCE contrastive loss formulation [8, 28, 45, 56] of the InfoMax principle [40] which maximizes the mutual information between embeddings of multiple views of x, x_i and x_j. In contrastive learning, the network is trained to correctly match each input sample with an augmented version of itself among a large training batch of other samples and their respective augmentations. We use the NT-Xent loss [8] defined as,

$$\mathcal{L}^{\text{NT-Xent}}(x_i, x_j) = -\log \frac{e^{sim(z_i,z_j)/\tau}}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} e^{sim(z_i,z_k)/\tau}} \quad (1)$$

where $sim(z_i, z_k)$ is the cosine similarity between z_i and z_k, τ is a temperature scalar and $z_k = g(x_k)$. N is the size of the mini-batch of distinct samples where each sample x has x_i and x_j as positive augmentations. As shown in Eqn. 1, the NT-Xent loss maximizes the agreement between two augmented views x_i and x_j of the same input sample x in a low-dimension representation space encoded by g.

For x_i^{act} and $x_{i1}^{ap}, \ldots, x_{iF}^{ap}$, the action and appearance stream encodings: h_i^{act} and $\bar{h}_i^{ap}$ are fed to MLP projection heads to obtain z_i^{act} and z_i^{ap}. Similarly we obtain z_j^{ap} and z_j^{act} for another augmentation set x_j^{act} and $x_{j1}^{ap}, \ldots, x_{jF}^{ap}$. z_i^{ap} and z_j^{ap} are used to compute $\mathcal{L}_{NCE}^{ap}$ contrastive loss to train the appearance encoder, while $\mathcal{L}_{NCE}^{act}$ is computed using z_i^{act} and z_j^{act} to train the action encoder.

3.3. Unsupervised Meta-learning for Video Few-Shot (MetaUVFS)

MetaUVFS explicitly trains a few-shot meta-learner via episodic training to improve performance on the downstream few-shot tasks having novel classes. MetaUVFS first generates episodes at video instance level using noise-contrastive embeddings without any supervision and imposes a hardness threshold to boost few-shot meta-learning. Using these generated episodes, MetaUVFS trains a novel Action-Appearance Aligned Meta-adaptation (A3M) module to align and relate action and appearance features, and output an embedding that can more effectively generalize to novel classes in few-shot testing. The episodic training of MetaUVFS imposes a downstream task-specific prior on the unsupervised model features that reduces the gap between train and test settings, thereby improving performance.

3.3.1 Unsupervised Hard Episodes Generation

To simulate the meta-testing episodic setting during training, we leverage the unlabeled video data to generate meaningful episodes for meta-training the A3M module. We generate 1-shot, 5-way classification episodes (similar to the downstream few-shot task) where the support and query for each class are formed using spatio-temporal augmentations (Sec. 4.2) of an unlabeled video sample. In this way, the classification happens at the instance level (*i.e.* each video behaves as its own class) and the task is to classify a query augmentation belonging to the correct video sample. A simple approach would be to randomly sample unlabeled videos and process their augmentations into episodes. However, the InfoNCE contrastive learning pushes the embeddings, h_i^{act} and $\bar{h}_i^{ap}$, for the augmentations of a video x_i already very close to each other compared to embeddings

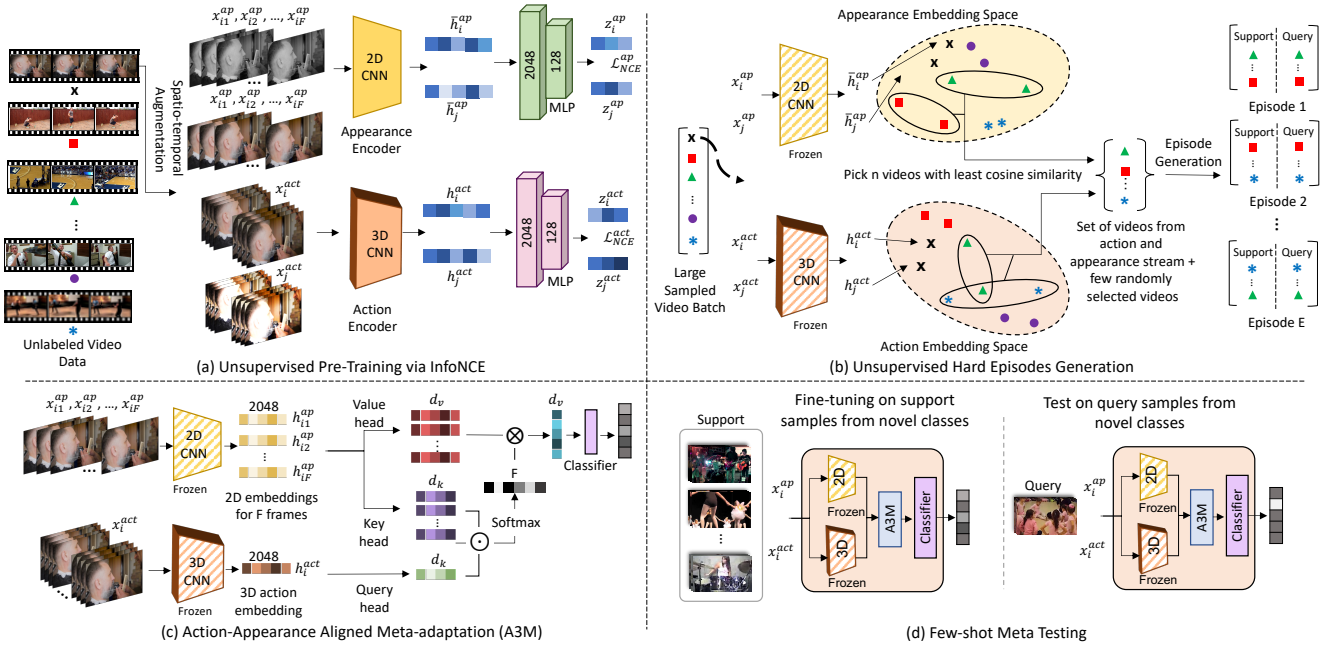

Figure 2. MetaUVFS: Model Overview. (a) **Two-stream Training**: using a large unlabeled video dataset subjected to our sampling and augmentation scheme (see Section 4.2). (b) **Hard Episode Generation** hard episodes are mined from the two-stream networks's feature space. (c) **A3M** module learns to align appearance over action features through episodic meta-adaptation. (d) **Meta-testing** Meta-trained A3M as a specialized few-shot classifier finetunes on novel class support to classify query videos.

for augmentations for other videos. Thus, randomly sampled videos will provide A3M module episodes that can be trivially solved and will impede any meaningful learning. As shown in Fig. 2b, to incentivize learning and generate meaningful episodes, we mine episodes where we select hard video instances whose augmentations lie far from each other in the embedding space of the trained two-stream encoders. We feedforward a large batch of video augmentations through the trained and frozen action and appearance encoders. For each encoder, we select n videos which have the lowest cosine similarity among its augmentations. We pool the set of the videos collected this way from the two encoders and extend this video set by another 10% with randomly sampled videos for exploration and to cover all video samples on expectation. We then sample E episodes from this set of videos for one training iteration of few-shot training. Selecting n videos for both action and appearance enables the few-shot meta-learner to reduce confusion from both action and appearance perspectives (Fig. 1).

3.3.2 A3M: Action-Appearance Aligned Meta-adaptation

As shown in Figure 1, it is important for the model to attend to both action and appearance-related aspects of a video in correspondence to each other to enhance intra-class relationship and avoid inter-class confusion, particularly when learning from very few labeled samples. To this end, we design a novel cross-attention module for action-appearance aligned meta-adaptation, A3M, that is trained using episodic few-shot learning to meta-learn to cross-align action with appearance-related features.

The A3M module learns to establish a soft correspondence between the action and appearance features using attention-based Transformers [60]. As shown in Fig. 2c, we parameterize three linear mappings, key-head $\mathbf{K} : \mathbb{R}^D \rightarrow \mathbb{R}^{d_k}$, value-head $\mathbf{V} : \mathbb{R}^D \rightarrow \mathbb{R}^{d_v}$ and query-head $\mathbf{Q} : \mathbb{R}^D \rightarrow \mathbb{R}^{d_k}$ for this purpose where d_k and d_v are the size of the key and value embeddings, respectively. We generate key-value pairs using $\mathbf{K}$ and $\mathbf{V}$ for the frame-level representations, $h^{ap}_{i1}, \ldots, h^{ap}_{iF}$, from the 2D appearance encoder. Let $\mathbf{k}_m = \mathbf{K} \cdot h^{ap}_m$ and $\mathbf{v}_m = \mathbf{V} \cdot h^{ap}_m$ form the key-value pair for the m^{th} frame-level representation for unlabeled x_i. We also generate a query embedding, $q = \mathbf{Q} \cdot h^{act}$, for the spatio-temporal feature, h^{act}, from the 3D action encoder using $\mathbf{Q}$. We then compute the dot-product attention scores between the keys and the query, and normalize the scores via softmax over all key embeddings as,

$$a_m = \frac{\exp(\mathbf{k}_m \cdot \mathbf{q})/\sqrt{d_k}}{\sum_t \exp(\mathbf{k}_t \cdot \mathbf{q}/\sqrt{d_k})} \tag{2}$$

where a_m is the attention score for the m^{th} frame embedding. These attention scores provide a soft correspondence that align and relate the action information with the appearance of the video. The attention scores are then combined with the value head embeddings and aggregated via sum to obtain a single feature embedding, $\mathbf{h}^{\mathbf{A3M}} = \sum_m a_m \mathbf{v}_m$. As the attention scores are computed via a combination of

action and appearance features, they weigh the appearance features to focus on the most action-relevant parts. The aggregated embedding $\mathbf{h}^{\mathbf{A3M}}$, conditioned on both action and appearance information, is therefore better equipped than naive concatenation for few-shot tasks.

3.3.3 Few-Shot Meta-Training

We leverage Model-Agnostic Meta Learning (MAML) [14] to train the network to learn to adapt to a new task of novel action classes with few labeled samples. Once we train the action and appearance streams, we freeze the two backbones and train f_θ comprising of the A3M module along with a classifier layer during the few-shot episodic meta-training. The action-appearance aligned feature embedding from the A3M module is l_2-normalized before being fed to the classifier. For each generated episode $e \in E$ in a training iteration, we generate s support augmentations for sampled videos and compute adapted parameters with gradient descent of the cross-entropy classification loss $\mathcal{L}$ over f_θ as $\theta'_e = \theta - \alpha \nabla_\theta \mathcal{L}_e(f_\theta)$ where α is the adaptation learning rate. We then generate q query augmentations for videos in episode e to compute the loss $\mathcal{L}$ using adapted parameters θ'_e as $\mathcal{L}_e(f_{\theta'_e})$. We repeat this for all E episodes and finally update θ at the end of the training iteration as $\theta \leftarrow \theta - \beta \nabla_\theta \sum_e^E \mathcal{L}_e(f_{\theta'_e})$ where β is the learning rate for the meta-learner optimizer.

3.3.4 Few-Shot Meta-Testing

Once trained, we test MetaUVFS by finetuning on multiple few-shot test episodes. As can be seen in Fig. 2d, for each episode, we freeze the action-appearance encoders and finetune the A3M and classifier layers which has been meta-trained. After every episode, we refresh the parameters of A3M and classifier layers for the next episode.

4. Experiments and Results

4.1. Datasets

We evaluate MetaUVFS on three publicly-available few-shot datasets: Kinetics100 [6, 74], UCF101 [52] and HMDB51 [37]. Following [74], we obtain the *few-shot* train/validation/test splits with 64/12/24 non-overlapping classes for Kinetics100. For UCF101 and HMDB51, we follow the *few-shot* split from [71]. UCF101 contains 100 classes split into 70/10/20 and HMDB51 contains 51 classes split as 31/10/10. The test splits of each dataset are used for novel class evaluation in the meta-testing phase. For the unsupervised training of MetaUVFS's two-stream networks, we leverage Kinetics700 [6] without using any labels. Kinetics700 is a large-scale video classification dataset that covers 700 human action classes including human-object and human-human interactions. To increase the size of our unlabeled training data, we also include the videos from the base-classes of Kinetics100, UCF101, and HMDB51, without the labels. Altogether, we obtain around 550K video clips with a duration of around 10s each (25 FPS). We take extra precaution to ensure that there is no video in the training dataset belonging to the union of all the novel classes across all three evaluation datasets. This is to ensure that our testing is truly on a disjoint set of unseen classes.

4.2. Implementation Details

Data Sampling and Augmentation We develop a spatio-temporal sampling protocol that is most optimal for the unsupervised two-stream training and A3M-based few-shot training/testing. For an input video, the 2D appearance stream encodes 8 input frames where 1 frame is randomly sampled from each of 8 segments equally-partitioned along the video length. With focus on spatial information, we use a higher frame resolution of 224×224. We refer to this as 8×1. For the 3D-action stream, with the goal of encoding fine-grained spatio-temporal action information across video segments, we sample 4 clips across 4 equidistant segments of the video to form a 16 frame input. To balance the spatio-temporal information, we use a lower frame resolution of 112×112. We refer to this as 4×4. We follow SimCLR's protocol for spatial augmentation [8]: a composition of Random crops, Random horizontal flips, Random Color Jitter, Random grayscale, Gaussian blur. The spatial augmentation is clip-wise consistent, *i.e.*, the random seed is fixed across all frames of a video augmentation [70, 22].

MetaUVFS Training We use ResNet50 [29] backbone to train the 2D appearance stream and its 3D counterpart, ResNet50-3D [26], for the 3D action stream. The dimension of z^{ap} and z^{act} obtained from the MLP projection head is 128 (similar to [8]). We first train the action and appearance streams individually using losses $\mathcal{L}^{ap}_{NCE}$ and $\mathcal{L}^{act}_{NCE}$ respectively. We use a batch size of 512 and train both models for 300 epochs on 64 NVIDIA P100 GPUs. Following [30, 21], we do a gradual learning rate (LR) warmup for 5 epochs followed by a half-period cosine learning rate decay with SGD optimizer and 0.9 momentum. With 0.001 per-gpu LR, we also linearly scale the LR to 0.064.

For hard-mining episodes, n is set to 32. We set $d_k = 128$ and $d_v = 2048$ for the A3M module. For MAML, we set $E = 10$, $\alpha = 0.001$ and $\beta = 10$. We train for $20,000$ iterations using Adam optimizer and cosine annealing [1] for a total of $200K$ unsupervised hard-mined episodes. For more details, please refer to the supplementary material.

Few-shot Evaluation We evaluate MetaUVFS on all three datasets based on 5-way, 1-shot and 5-way, 5-shot settings as is standard in few-shot learning literature. For each episode, 5 classes are randomly sampled from the set

Methods	Supervision		UCF101		HMDB51		Kinetics100	
	Pretraining	Base-Class	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Matching Net [74]	Imagenet-2D	Yes	-	-	-	-	53.3	74.6
MAML [74]	Imagenet-2D	Yes	-	-	-	-	54.2	75.3
CMN [74]	Imagenet-2D	Yes	-	-	-	-	60.5	78.9
TARN [3]	Sports-1M	Yes	-	-	-	-	66.6	80.7
OTAM [5]	Imagenet-2D	Yes	-	-	-	-	73.0	85.8
R-3DFSV [68]	Sports-1M	Yes	-	-	-	-	95.3	97.8
ProtoGAN [38]	Sports-1M	Yes	57.8 ± 3.0	80.2 ± 1.3	34.7 ± 9.20	54.0 ± 3.90	-	-
AmeFu-Net [15]	Imagenet-2D	Yes	85.1	95.5	60.2	75.5	74.1	86.8
RVN [4]	Kinetics-400	Yes	88.71 ± 0.19	96.78 ± 0.08	63.43 ± 0.28	79.69 ± 0.20	-	-
ARN [71]	No	Yes	66.32 ± 0.99	83.12 ± 0.70	45.15 ± 0.96	60.56 ± 0.86	63.7	82.4
3DRotNet [33]	No	No	39.43 ± 0.48	33.61 ± 0.34	32.35 ± 0.42	27.84 ± 0.40	27.53 ± 0.36	25.54 ± 0.39
VCOP [70]	No	No	32.91 ± 0.42	39.11 ± 0.37	27.80 ± 0.37	31.56 ± 0.35	26.48 ± 0.37	28.87 ± 0.36
IIC [55]	No	No	56.81 ± 0.46	78.74 ± 0.37	34.66 ± 0.41	49.57 ± 0.44	37.73 ± 0.43	51.11 ± 0.43
Pace Prediction [62]	No	No	25.58 ± 0.33	26.58 ± 0.31	26.21 ± 0.33	27.09 ± 0.31	22.42 ± 0.33	22.94 ± 0.30
MemDPC [23]	No	No	49.27 ± 0.44	67.38 ± 0.45	30.33 ± 0.40	41.15 ± 0.42	42.01 ± 0.41	53.90 ± 0.43
CoCLR [24]	No	No	51.99 ± 0.46	72.17 ± 0.42	31.29 ± 0.40	44.92 ± 0.45	37.59 ± 0.42	51.11 ± 0.43
CVRL [47]	No	No	63.00 ± 0.41	87.80 ± 0.30	44.21 ± 0.45	60.35 ± 0.45	53.26 ± 0.48	71.39 ± 0.44
MetaUVFS (Ours)	No	No	**76.38 ± 0.40**	**92.50 ± 0.24**	**47.55 ± 0.45**	**66.13 ± 0.33**	**62.80 ± 0.45**	**79.55 ± 0.39**

Table 1. Results on UCF101, HMDB51 and Kinetics100 datasets for 5-way, 1-shot and 5-shot few-shot action recogniton. Our method MetaUVFS outperforms SoTA methods on unsupervised video representations by large margins on few-shot benchmarks. We also show competitive performance w.r.t. supervised few-shot video approaches. Moreover, on UCF101 and HMDB51, MetaUVFS is able to outperform ARN that uses only base-class supervision. MetaUVFS also outperforms ProtoGAN on UCF101 and HMDB51, and CMN on Kinetics100. Values in blue represent SoTA across all levels of supervision.

of novel classes for classification and training happens on 1 and 5 support samples per class respectively. In all settings, Top-1 accuracy is reported on 1 query sample per class. In each experiment, we randomly sample 10,000 episodes for few-shot meta-testing and report the average accuracy at the 95% confidence interval. Finetuning is done at a constant learning rate of 10 for 50 epochs for all experiments.

4.3. Compare to SoTA Unsupervised Approaches

Table 1 compares MetaUVFS with various state-of-the-art supervised and unsupervised methods on different few-shot settings and datasets. We categorize the different techniques based on the amount of supervision in terms of base-class data ('Yes' in *Base-Class*) and surrogate supervision, *i.e.*, initializing the network using the weights pretrained on a large-scale supervised image/video data ('Yes' in *Pretrained Weights*). Cells are left blank if there are no publicly available results for that setting.

The second part of Table 1 compares MetaUVFS with various state-of-the-art methods that leverage unlabeled videos for representation learning. To the best of our knowledge, MetaUVFS is the first approach that specializes in few-shot action recognition in a purely unsupervised manner. Hence, there is no publicly available benchmark for the performance of existing video-based unsupervised techniques on few-shot action recognition. We took the initiative to assess these approaches on our few-shot test-bed using the same hyperparameters for few-shot meta-testing as MetaUVFS. Many of these approaches are originally trained on a relatively small unlabeled dataset. Therefore, for a fair comparison, we train these methods on our large-scale unlabeled dataset using their publicly available code.

As shown in Table 1, MetaUVFS is able to clearly outperform all state-of-the-art unsupervised methods on the task of few-shot action recognition by at least 13.38%, 3.34% and 9.54% (absolute increase) on UCF101, HMDB51 and Kinetics100 1-shot, 5-way benchmark respectively. Among the methods we compare, IIC [55], CVRL [47] and CoCLR (RGB only) [24] also use contrastive loss for unsupervised training. The superior performance of MetaUVFS in comparison to these methods indicate that our approach of jointly leveraging and aligning action and appearance along with meta-training episodically for few-shot plays an integral role in performing effectively when the downstream task lies in the low-shot regime.

4.4. Compare to SoTA Supervised Few-shot Works

The first part of Table 1 compares MetaUVFS with various state-of-the-art supervised few-shot action recognition methods. We can observe that compared to ARN [71] that uses only base-class data as supervision, MetaUVFS significantly outperforms on UCF101 and HMDB51, and performs competitively on Kinetics100. Furthermore, MetaUVFS is even able to outperform some of the supervised methods that use both pretrained weights and base-class labels for supervision such as ProtoGAN [38] on UCF101 and HMDB51, and CMN [74] on Kinetics100. It is worth noting that, unlike these methods that need to train separate models to obtain results on the different datasets, MetaUVFS trains a *single* unsupervised model to achieve all results. This single model either outperforms or performs competitively compared to supervised methods across all three datasets.

Action	Appearance	A3M	Hard Episodes	UCF101		HMDB51		Kinetics100	
				1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
✓			✓	66.97 ± 0.44	88.64 ± 0.30	44.56 ± 0.45	61.03 ± 0.45	53.56 ± 0.48	71.31 ± 0.44
	✓		✓	66.10 ± 0.45	84.58 ± 0.34	39.97 ± 0.46	58.20 ± 0.46	50.91 ± 0.48	69.47 ± 0.46
✓	✓			71.82 ± 0.42	89.93 ± 0.27	44.64 ± 0.46	62.40 ± 0.45	57.52 ± 0.46	75.21 ± 0.41
✓	✓		✓	73.02 ± 0.42	91.38 ± 0.26	44.89 ± 0.46	64.96 ± 0.44	59.16 ± 0.44	77.42 ± 0.40
✓	✓	✓		73.97 ± 0.41	91.50 ± 0.26	45.84 ± 0.45	64.68 ± 0.41	59.88 ± 0.45	77.64 ± 0.39
✓	✓	✓	✓	**76.38 ± 0.40**	**92.50 ± 0.24**	**47.55 ± 0.45**	**66.13 ± 0.33**	**62.80 ± 0.45**	**79.55 ± 0.39**

Table 2. Ablation study of MetaUVFS highlights the superior few-shot performance of meta-training two-stream feature representations over individual action and appearance streams. The performance is further boosted by action-appearance feature alignment by the A3M module. Moreover, mining unsupervised hard episodes is crucial for effectively training the A3M module.

4.5. MetaUVFS: Ablation Study

We conduct an ablation study where we isolate individual aspects of MetaUVFS and quantify their impact on the few-shot performance. Table 2 summarizes the results. We train all ablation experiments using MAML as the few-shot algorithm.

We first conduct experiments without the A3M module where the network consists of only the action stream, only the appearance stream and dual action-appearance stream (Table 2, Rows 1, 2, 4). Without the A3M module, for the one-stream setting, we directly feed the features from the available stream (averaging appearance features over 8 frames) to the classifier layer for few-shot episodic meta-training and later for meta-testing; for the two-stream setting, we simply concatenate the action features and appearance features (averaged over 8 frames) and feed them to the classifier for few-shot episodic meta-training. All three experiments use unsupervised hard-mined episodes. In the absence of either the action or the appearance stream, only the features of the available stream are used to mine episodes. We can observe from Table 2 (Rows 1, 2, 4) that the few-shot performance is significantly worse when either action or appearance stream is missing compared to when both are present. This is because when only a few support samples are available to learn for a set of novel classes, the likelihood of the model to make mistakes reduces sharply in the presence of both streams as it allows the network more ways to activate and respond to the representative features necessary for correct classification. We can also compare Row 1 (Action stream only) with CVRL [47] in Table 1. CVRL backbone is similar to our 3D action stream. However, due to an explicit few-shot training phase, our *Action only* baseline performs consistently better than CVRL.

Rows 4 and 6 in Table 2 highlight the impact of the A3M module by aligning action-appearance as part of few-shot training. Our proposed A3M module in MetaUVFS results in an average absolute improvement of 3.22% and 1.47% on 5-way, 1-shot and 5-way, 5-shot benchmarks across all datasets. Aligning the action and appearance features during few-shot episodic training significantly improves the model's ability to *attend* to the most representative video aspects while leveraging the inductive biases of both 2D and 3D CNNs to learn complementary representation that boosts few-shot performance.

We then perform an ablation where we train our method episodically without mining hard episodes based on noise-contrastive embeddings (Table 2, Row 5). Comparing Rows 5 and 6, we can observe a significant reduction in performance without hard episodes, underlining the importance of mining hard episodes to the few-shot episodic training of MetaUVFS. This is because in the absence of hard episodes, the randomly sampled videos in a training episode are such that the action and appearance embeddings fed for support and query augmentation samples to the A3M module during training are already easily separable. This severely compromises the training of A3M and makes it behave close to an identity function, as evident from Row 5's only slightly higher performance than Row 4 where A3M is not present.

We additionally perform an experiment where both A3M and hard episodes are absent during training (Table 2, Row 3). We can observe that this setting results in a statistically significant reduction in performance compared to when A3M and/or hard episodes are employed for training (Rows 4-6).

4.6. Discussion

Impact of Frame Sampling. Since the two-streams in MetaUVFS specialize both in terms of architecture and their function, we observe that the sampling strategy in choosing the frames as input to both streams along with their frame resolution make a difference in the performance. Table 3 provides an analysis of the few-shot performance for 5-way, 1-shot settings on Kinetics100 across different sampling strategies for both 3D action and 2D appearance streams first individually and then in combination. We observe that for the 3D stream, choosing a 4×4 sampling, *i.e.*, sampling 4 segments of 4 frames uniformly over the entire video length provides a 3.3% improvement over sampling 16 frames from 32 consecutive frames with a stride of 2. Similarly, for the 2D stream, we find that 16×1 and 8×1 sampling, *i.e.*, sampling 1 frame from 16 or 8 segments over the entire video length as most effective. In our two-stream setting, we find 8×1, 4×4 as the optimal sampling scheme.

Meta-learning Algorithm. We further validate the choice of MAML as our few-shot meta-learning algorithm by as-

	Appearance only (224×224)			Action only (112×112)		
Sampling	4×1	8×1	16×1	32→16	4×4	8→4×4
1-shot	50.54 ± 0.48	50.91 ± 0.48	51.0 ± 0.42	50.26 ±0.52	53.56 ± 0.48	53.09 ± 0.42
	MetaUVFS(Appearance + Action + A3M, ours)					
Sampling	4×1, 4×4	4×1, 8→4×4	8×1, 4×4	8×1, 8→4×4	16×1, 4×4	16×1, 8→4×4
1-shot	60.76 ± 0.44	60.69 ± 0.45	**62.80 ± 0.45**	61.34 ±0.48	60.30 ±0.41	60.91 ± 0.47

Table 3. Comparing sampling strategies evaluated on Kinetics100 for 1-shot, 5-way. → denotes downsampling from A to B value.

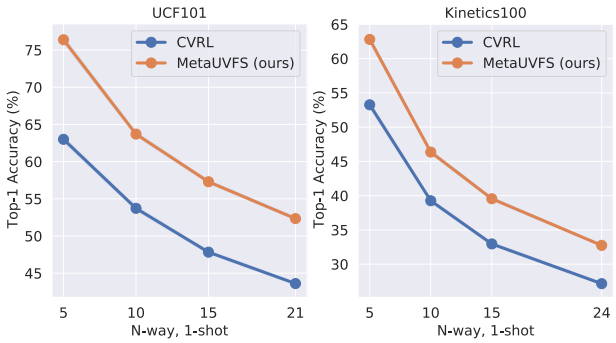

Figure 3. Comparison of MetaUVFS with CVRL [47] on UCF101 and Kinetics100 dataset on 1-shot few-shot for N-way classification where N ranges from 5 to over 20.

Few-shot Algorithm	1-shot	5-shot
ProtoNet [51]	31.20 ± 0.39	55.96 ± 0.44
Baseline++ [9]	56.10 ± 0.47	73.37 ± 0.43
ProtoMAML [59]	62.13 ± 0.46	78.65 ± 0.40
MAML [14]	**62.80 ± 0.45**	**79.55 ± 0.39**

Table 4. Comparison between different few-shot meta-learning algorithms on Kinetics100 5-way, 1/5-shot dataset.

sessing other popular few-shot approaches in the literature. To compare with ProtoNet [51] and ProtoMAML [59], we repurpose the output of A3M module to compute prototypes across support samples (augmentations) and compare against query samples to compute the loss during training. For few-shot testing on novel classes, ProtoNet matches the query samples with prototypes that are computed from support samples to assign class label based on the best matched prototype. For ProtoMAML, we reshape the prototypes computed from support samples as parameters of the classifier layer which is finetuned along with the A3M module as per Sec. 3.3.4. We also compare with Baseline++ [9] where we use meta-trained parameters for A3M and classifier but change the few-shot finetuning during testing to Baseline++. As shown in Table 4, we find that our few-shot meta-testing protocol of using MAML with l_2-normalized embedding as input to the classifier significantly outperforms other few-shot learning methods. We also observe that, in general, few-shot methods employing finetuning during few-shot testing tend to perform better. We believe this is due to extra adaptation steps needed during testing because of the absence of supervision during training.

Significance of Action-Appearance We conduct an experiment where both the streams are either 3D CNNs (action) or 2D CNNs (appearance). This delineates the impact of having complementary action and appearance streams on few-shot performance from the impact of increase in the parameter count due to an additional backbone. We train them using MAML with hard episodes without A3M for fair comparison. We can observe from Table 5 that although having more parameters with either two appearance or two action streams improves the performance compared to single stream, the improvement is significantly higher when a combination of action and appearance streams is used that helps to leverage more diverse 2D/3D representations to learn more generalizable few-shot video representations.

Streams	1-shot	5-shot
Appearance + Appearance	55.75 ± 0.46	72.23 ± 0.43
Action + Action	54.25 ± 0.47	73.21 ± 0.42
Action + Appearance	**59.16 ± 0.44**	**77.42 ± 0.40**

Table 5. Comparison of using Action and Appearance streams in MetaUVFS with using two Action or two Appearance streams. Results are without A3M module for fair comparison.

Many-Way Few Shot. We go beyond the 5-way 1-shot setting by increasing the number of novel classes to evaluate MetaUVFS on a more challenging and *in-the-wild* few-shot many-way classification task on UCF101 and Kinetics100. Fig. 3 shows a plot for this experiment. Although, as expected, the performance reduces with increasing N, MetaUVFS is still able to outperform the best-performing unsupervised baseline method, CVRL, by a significant margin for all N-way 1-shot classification tasks considered. This proves that MetaUVFS is more robust even in extreme few-shot scenarios where the inter-class confusion is higher.

5. Conclusion

We propose a novel unsupervised meta-learning algorithm, MetaUVFS, for few-shot video action recognition. It leverages large-scale unlabeled video data to learn unsupervised video features from a two-stream action-appearance network. It further performs explicit few-shot episodic meta-learning over unsupervised hard-mined episodes using a novel Action-Appearance Aligned Meta-adaptation (A3M) module. The A3M module learns to align the 3D action with 2D appearance features to learn an embedding that is more effective in focusing on the action-specific features of a video for the few-shot downstream task. Through extensive experiments, we demonstrate that using an explicit few-shot learner and action-appearance aligned features makes MetaUVFS significantly better suited for downstream few-shot tasks compared to all state-of-the-art unsupervised methods. Moreover, MetaUVFS performs competitively and sometimes even outperforms SoTA supervised few-shot methods.

References

[1] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. 2019. 5

[2] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9922–9931, 2020. 3

[3] Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. *arXiv preprint arXiv:1907.09021*, 2019. 2, 6

[4] Congqi Cao, Yajuan Li, Qinyi Lv, Peng Wang, and Yanning Zhang. Few-shot action recognition with implicit temporal alignment and pair similarity optimization. *arXiv preprint arXiv:2010.06215*, 2020. 1, 2, 6

[5] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 6

[6] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 5

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 3, 5

[9] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2018. 8

[10] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019. 1, 2

[11] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015. 3

[12] Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: spatially-aware few-shot transfer. *arXiv preprint arXiv:2007.11498*, 2020. 2

[13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 6202–6211, 2019. 2

[14] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017. 1, 2, 5, 8

[15] Yuqian Fu, Li Zhang, Junke Wang, Yanwei Fu, and Yu-Gang Jiang. Depth guided adaptive meta-fusion network for few-shot video recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1142–1151, 2020. 1, 2, 6

[16] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8059–8068, 2019. 2

[17] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4367–4375, 2018. 1, 2

[18] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 3

[19] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 971–980, 2017. 2

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27:2672–2680, 2014. 2

[21] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 5

[22] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. *arXiv preprint arXiv:2008.01065*, 2020. 3, 5

[23] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *ECCV*, 2020. 6

[24] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *Neurips*, 2020. 1, 3, 6

[25] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3154–3160, 2017. 2

[26] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018. 5

[27] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3018–3027, 2017. 1, 2

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[30] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 558–567, 2019. 5

[31] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. *arXiv preprint arXiv:1910.07677*, 2019. 2

[32] Kyle Hsu, Sergey Levine, and Chelsea Finn. Unsupervised learning via meta-learning. In *International Conference on Learning Representations*, 2018. 3

[33] Longlong Jing, Xiaodong Yang, Jingen Liu, and Yingli Tian. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv preprint arXiv:1811.11387*, 2018. 3, 6

[34] Siavash Khodadadeh, Ladislau Boloni, and Mubarak Shah. Unsupervised meta-learning for few-shot image classification. In *Advances in Neural Information Processing Systems*, pages 10132–10142, 2019. 3

[35] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8545–8552, 2019. 3

[36] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015. 1

[37] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011. 5

[38] Sai Kumar Dwivedi, Vikram Gupta, Rahul Mitra, Shuaib Ahmed, and Arjun Jain. Protogan: Towards few shot learning for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2, 6

[39] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017. 3

[40] Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988. 3

[41] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016. 3

[42] Carlos Medina, Arnout Devos, and Matthias Grossglauser. Self-supervised prototypical transfer learning for few-shot classification. *arXiv preprint arXiv:2006.11325*, 2020. 3

[43] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016. 3

[44] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016. 3

[45] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[46] AJ Piergiovanni, Anelia Angelova, and Michael S Ryoo. Evolving losses for unsupervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 133–142, 2020. 3

[47] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. *arXiv preprint arXiv:2008.03800*, 2020. 1, 3, 6, 7, 8

[48] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016. 1, 2

[49] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018. 2

[50] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850. PMLR, 2016. 2

[51] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017. 1, 2, 8

[52] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[53] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018. 1, 2

[54] Li Tao, Xueting Wang, and Toshihiko Yamasaki. Rethinking motion representation: Residual frames with 3d convnets for better action recognition. *arXiv preprint arXiv:2001.05661*, 2020. 2

[55] Li Tao, Xueting Wang, and Toshihiko Yamasaki. Self-supervised video representation learning using inter-intra contrastive framework. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2193–2201, 2020. 1, 3, 6

[56] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 3

[57] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

[58] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 2

[59] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *International Conference on Learning Representations*, 2019. 8

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008, 2017. 4

[61] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016. 1, 2

[62] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. *arXiv preprint arXiv:2008.05861*, 2020. 3, 6

[63] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 2

[64] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 2

[65] Yan Wang, Wei-Lun Chao, Kilian Q Weinberger, and Laurens van der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019. 2

[66] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7278–7286, 2018. 1, 2

[67] Chao-Yuan Wu, Ross Girshick, Kaiming He, Christoph Feichtenhofer, and Philipp Krahenbuhl. A multigrid method for efficiently training video models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 153–162, 2020. 2

[68] Yongqin Xian, Bruno Korbar, Matthijs Douze, Bernt Schiele, Zeynep Akata, and Lorenzo Torresani. Generalized many-way few-shot video classification. *arXiv preprint arXiv:2007.04755*, 2020. 1, 2, 6

[69] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321, 2018. 2

[70] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019. 3, 5, 6

[71] Hongguang Zhang, Li Zhang, X Qui, Hongdong Li, Philip HS Torr, and Piotr Koniusz. Few-shot action recognition with permutation-invariant attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 5, 6

[72] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 3

[73] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 2

[74] Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 751–766, 2018. 1, 2, 5, 6