Inductive and Transductive Few-Shot Video Classification via Appearance and Temporal Alignments

Khoi D. Nguyen[1], Quoc-Huy Tran[2], Khoi Nguyen[1], Binh-Son Hua[1], and Rang Nguyen[1]*

[1] VinAI Research, Vietnam
[2] Retrocausal, Inc., USA

Abstract. We present a novel method for few-shot video classification, which performs appearance and temporal alignments. In particular, given a pair of query and support videos, we conduct appearance alignment via frame-level feature matching to achieve the appearance similarity score between the videos, while utilizing temporal order-preserving priors for obtaining the temporal similarity score between the videos. Moreover, we introduce a few-shot video classification framework that leverages the above appearance and temporal similarity scores across multiple steps, namely prototype-based training and testing as well as inductive and transductive prototype refinement. To the best of our knowledge, our work is the first to explore transductive few-shot video classification. Extensive experiments on both Kinetics and Something-Something V2 datasets show that both appearance and temporal alignments are crucial for datasets with temporal order sensitivity such as Something-Something V2. Our approach achieves similar or better results than previous methods on both datasets. Our code is available at https://github.com/VinAIResearch/fsvc-ata.

Keywords: Few-Shot Learning, Video Classification, Appearance Alignment, Temporal Alignment, Inductive Inference, Transductive Inference

1 Introduction

Recognizing video contents plays an important role in many real-world applications such as video surveillance [51,34], anomaly detection [41,17], video retrieval [13,39], and action segmentation [22,21]. In the modern era of deep learning, there exist a large number of studies focusing on learning to classify videos by fully supervising a neural network with a significant amount of labeled data [43,6,47,44]. While these fully-supervised approaches provide satisfactory results, the high costs of data collection and annotation make it unrealistic to transfer an existing network to new tasks. To reduce such high costs, few-shot learning [37,45,38,11,25,32,42,36,31] is an emerging trend that aims to

* Corresponding author

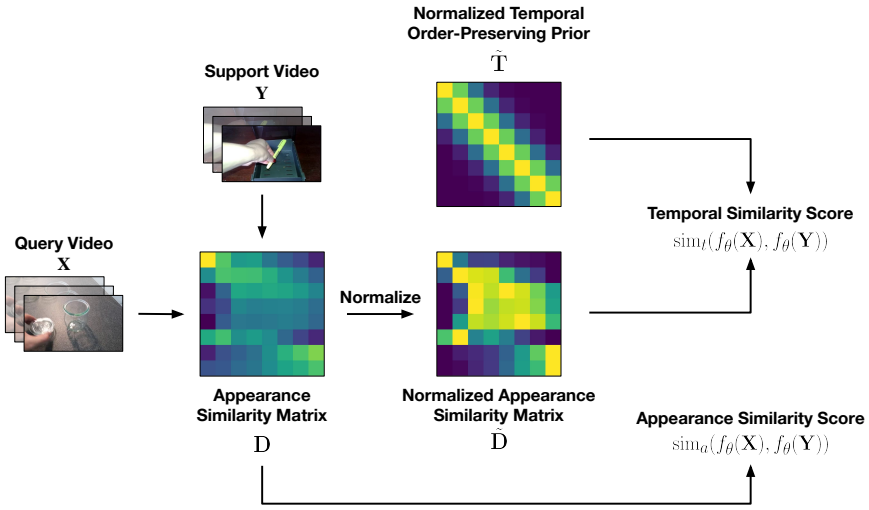

Fig. 1. Our Few-Shot Video Classification Approach. Given a pair of query and support videos, we first perform appearance alignment (i.e., via frame-level feature matching) to compute the appearance similarity score, and then temporal alignment (i.e., by leveraging temporal order-preserving priors) to calculate the temporal similarity score. The final alignment score between the two videos is the weighted sum of appearance and temporal similarity scores.

adapt an existing network to recognize new classes with limited training data. Considerable research efforts have been invested in few-shot learning on images [46,10,24,50,9,29,14]. Extending few-shot learning to videos has been rather limited.

The main difference between a video and an image is the addition of temporal information in video frames. To adapt few-shot learning for videos, recent emphasis has been put on temporal modeling that allows the estimate of the similarity between two videos via frame-to-frame alignment. In the few-shot setting, this similarity function is crucial because it helps classify a query video by aligning it with the given support videos. Early methods [53,2,12,1,4] achieve low performance as they neglect temporal modeling and simply collapse the temporal information in their video representation learning. Recent methods [5,28] jointly consider appearance and temporal information by using Dynamic Time Warping [8] or Optimal Transport [7] to align the videos.

We propose to separately consider appearance and temporal alignments, yielding robust similarity functions for use in training and testing in both inductive and transductive few-shot learning. Following previous works [5,57], we sparsely sample a fixed number of frames from each video and extract their corresponding features using a neural network-based feature extractor. We then compute the pairwise cosine similarity between frame features of the two videos, yielding the appearance similarity matrix. To compute the appearance similar-

ity score between the videos, we first match each frame from one video to the most similar frame in the other, ignoring their temporal order. We then define our appearance similarity score between the two videos as the total similarity of all matched frames between them. Next, motivated by the use of temporal order-preserving priors in different video understanding tasks [40,5,16,23] (i.e., initial frames from a source video should be mapped to initial frames in a target video, and similarly the subsequent frames from the source and target video should match accordingly), we encourage the appearance similarity matrix to be as similar as possible to the temporal order-preserving matrix. Our temporal similarity score between the two videos is then computed as the negative Kullback-Leibler divergence between the above matrices. Furthermore, we show how to apply the above appearance and temporal similarity scores in different stages of few-shot video classification, from the prototype-based training and testing procedures to the finetuning of prototypes in inductive and transductive settings. Our method achieves the state-of-the-art performance on both Kinetics [20] and Something-Something V2 [15] datasets. Fig. 1 illustrates our ideas.

In summary, our contributions include:

- We introduce a novel approach for few-shot video classification leveraging appearance and temporal alignments. Our main contribution includes an appearance similarity score based on frame-level feature matching and a temporal similarity score utilizing temporal order-preserving priors.
- We incorporate the above appearance and temporal similarity scores into various steps of few-shot video classification, from the prototype-based training and testing procedures to the refinement of prototypes in inductive and transductive setups. To our best knowledge, our work is the first to explore transductive few-show video classification.
- Extensive evaluations demonstrate that our approach performs on par with or better than previous methods in few-shot video classification on both Kinetics and Something-Something V2 datasets.

2 Related Work

Few-Shot Learning. Few-shot learning aims to extract task-level knowledge from seen data while effectively generalizing learned meta-knowledge to unknown tasks. Based on the type of meta-knowledge they capture, few-shot learning techniques can be divided into three groups. Memory-based methods [37,30] attempt to solve few-shot learning by utilizing external memory. Metric-based methods [45,38,32,42,18,49,19,52] learn an embedding space such that samples of the same class are mapped to nearby points whereas samples of different classes are mapped to far away points in the embedding space. Zhang et al. [52] introduce the earth mover's distance to few-shot learning, which performs spatial alignment between images. Optimization-based approaches [11,25,36,31,55] train a model to quickly adapt to a new task with a few gradient descent iterations. One notable work is MAML [11], which learns an initial representation that can effectively be finetuned for a new task with limited labelled data. A new group of

works [24,50,9] utilize information from meta-training dataset explicitly during meta-testing to boost the performance. While Le et al. [24] and Das et al. [9] use base samples as distractors for refining classifiers, Yang et al. [50] finds similar features from base data to augment the support set.

Few-Shot Video Classification. Zhu and Yang [56] extend the notion of memory-based meta-learning for video categorization by learning many key-value pairs to represent a video. However, the majority of existing few-shot video classification methods are based on metric-based meta-learning, in which a generalizable video metric is learned from seen tasks and applied to novel unseen tasks. The main difference between a video and an image is the extra temporal dimension of the video frames. Extending few-shot learning to videos thus requires the temporal modeling of this extra dimension. This field of research can be divided in two main groups: aggregation-based and matching-based. The former prioritizes semantic contents, whereas latter is concerned with temporal orderings. Both of these groups start by sampling frames or segments from a video and obtaining their representations using a pretrained encoder. However, aggregation-based methods generate a video-level representation for distance calculation via pooling [1,4], adaptive fusion [2,12], or attention [53], whereas matching-based methods explicitly align two sequences and define distance between the videos as their matching cost via different matching techniques such as Dynamic Time Warping (DTW) [5] and Optimal Transport (OT) [28]. While [5] performs temporal alignment between videos via DTW, which strictly enforces a monotonic frame sequence ordering, [28] handles permutations using OT with the iterative Sinkhorn-Knopp algorithm, which can be practically slow and poorly scaled [48]. Our proposed method balances between appearance and temporal alignments which allows permutations to some extent. Moreover, it is non-iterative, hence more computationally efficient. Lastly, [5] and [28] explore the inductive setting only.

Transductive Inference. Transductive learning approaches exploit both the labeled training data and the unlabelled testing data to boost the performance. Many previous works [24,3,18,58,33] explicitly utilize unsupervised information from the query set to augment the supervised information from the support set. With access to more data, transductive inference methods typically outperform their inductive counterparts. However, existing transductive methods are designed mostly for few-shot image classification. In this work, we develop a cluster-based transductive learning approach with a novel assignment function leveraging both temporal and appearance information to address few-shot video classification.

3 Our Method

We present, in this section, the details of our approach. We first describe the problem of few-shot video classification in Sec. 3.1. Next, we introduce our appearance and temporal similarity scores in Sec. 3.2, which are then used in our prototype-based training and testing presented in Sec. 3.3. Lastly, we discuss

how the prototypes are refined in both inductive and transduction settings in Sec. 3.4.

3.1 Problem Formulation

In few-shot video classification, we are given a base set $\mathcal{D}_b = \{(\mathbf{X}_i, y_i)\}_{i=1}^{T_b}$, where $\mathbf{X}_i$ and y_i denote a video sample and its corresponding class label for N_b classes, respectively, while T_b is the number of samples. This base set is used for training a neural network which is subsequently adapted to categorize unseen videos with novel classes. At test time, we are given a set of support videos $\mathcal{D}_n^s = \{(\mathbf{X}_i, y_i)\}_{i=1}^{N \times K}$, where N and K denote the number of novel classes and the number of video samples per novel class, respectively. Note that the novel classes do not overlap with the base classes. The support set provides a limited amount of data to guide the knowledge transfer from the base classes to the novel classes. The goal at test time is to classify the query videos $\mathcal{D}_n^q$ into one of these novel classes. Such configuration is called an N-way K-shot video classification task. In this paper, we explore two configurations: 5-way 1-shot and 5-way 5-shot video classification.

There are two settings for few-shot learning: inductive and transductive learning. In the former, each of the query videos are classified independently, whereas, in the latter, the query videos are classified collectively, allowing unlabeled visual cues to be shared and leveraged among the query videos, which potentially improves the overall classification results. In this work, we consider both inductive and transductive settings. To our best knowledge, our work is the first to explore transductive learning for few-shot video classification.

Following prior works [5,57], we represent a video by a fixed number of M frames randomly sampled from equally separated M video segments, or $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^M]$ with $\mathbf{x}^i \in \mathbb{R}^{3 \times H \times W}$ and $i \in \{1, \ldots, M\}$, while H and W are the width and height of the video frame, respectively. Next, a neural network f with parameters θ is used to extract a feature vector $f_\theta(\mathbf{x}^i) \in \mathbb{R}^C$ (C is the number of channels) for each sampled frame $\mathbf{x}^i$. Lastly, the features of a video $\mathbf{X}$ is represented by $f_\theta(\mathbf{X}) = [f_\theta(\mathbf{x}^1), \ldots, f_\theta(\mathbf{x}^M)] \in \mathbb{R}^{C \times M}$.

3.2 Appearance and Temporal Similarity Scores

Existing distance/similarity functions are either computationally inefficient [28], imposing too strong constraint [5], or oversimplified that they neglect temporal information [57]. To capture both appearance and temporal information with low computational costs, we propose to explore appearance and temporal cues via two simple yet novel similarity functions. The final prediction is then a linear combination of the predictions from the two functions. We detailed our similarity functions, below.

Appearance Similarity Score. Given a pair of videos $(\mathbf{X}, \mathbf{Y})$, we first compute the appearance similarity matrix $\mathbf{D} \in \mathbb{R}^{M \times M}$ between them. The element $\mathbf{D}(i, j)$ of $\mathbf{D}$ is the pairwise cosine similarity between frame $\mathbf{x}^i$ in $\mathbf{X}$ with frame $\mathbf{y}^j$ in $\mathbf{Y}$

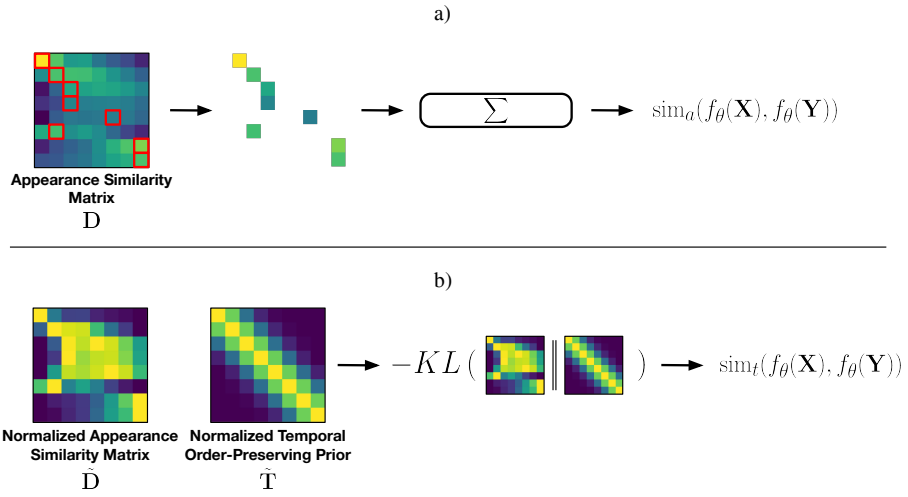

Fig. 2. Appearance and Temporal Similarity Scores. a) The proposed appearance similarity score is computed as the sum of the frame-level maximum appearance similarity scores between frames in $\mathbf{X}$ and frames in $\mathbf{Y}$. b) The proposed temporal similarity score is based on the Kullback-Leibler divergence between the row-wise normalized appearance similarity matrix $\tilde{\mathbf{D}}$ and the row-wise normalized temporal order-preserving prior $\tilde{\mathbf{T}}$.

as follows:

$$\mathbf{D}(i,j) = \frac{f_\theta(\mathbf{x}^i)^T f_\theta(\mathbf{y}^j)}{||f_\theta(\mathbf{x}^i)||\ ||f_\theta(\mathbf{y}^j)||}. \tag{1}$$

For every frame $\mathbf{x}^i$ in $\mathbf{X}$, we can align it with the frame $\mathbf{y}^k$ in $\mathbf{Y}$ which has the highest appearance similarity score with $\mathbf{x}^i$ (i.e., $k = \text{argmax}_j\, \mathbf{D}(i,j)$), ignoring their relative ordering. We define the appearance similarity score between $\mathbf{X}$ and $\mathbf{Y}$ as the sum of the optimal appearance similarity scores for all frames in $\mathbf{X}$ as:

$$\text{sim}_a(f_\theta(\mathbf{X}), f_\theta(\mathbf{Y})) = \sum_{i=1}^{M} \max_j \mathbf{D}(i,j) \approx \sum_{i=1}^{M} \lambda \log \sum_{j=1}^{M} \exp^{\frac{\mathbf{D}(i,j)}{\lambda}}, \tag{2}$$

where we use *log-sum-exp* (with the smoothing temperature $\lambda = 0.1$) to continuously approximate the *max* operator. Intuitively, the $\text{sim}_a(f_\theta(\mathbf{X}), f_\theta(\mathbf{Y}))$ shows how similar in appearance frames in $\mathbf{X}$ to frames in $\mathbf{Y}$. Note that sim_a is not a symmetric function. Fig. 2(a) shows the steps to compute our appearance similarity score.

Temporal Similarity Score. Temporal order-preserving priors have been employed in various video understanding tasks such as sequence matching [40], few-shot video classification [5], video alignment [16], and activity segmentation [23]. In particular, given two videos $\mathbf{X}$ and $\mathbf{Y}$ of the same class, it encourages initial

frames in $\mathbf{X}$ to be aligned with initial frames in $\mathbf{Y}$, while subsequent frames in $\mathbf{X}$ are encouraged to be aligned with subsequent frames in $\mathbf{Y}$. Mathematically, it can be modeled by a 2D distribution $\mathbf{T} \in \mathbb{R}^{M \times M}$, whose marginal distribution along any line perpendicular to the diagonal is a Gaussian distribution centered at the intersection on the diagonal, as:

$$\mathbf{T}(i,j) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{l^2(i,j)}{2\sigma^2}}, \quad l(i,j) = \frac{|i-j|}{\sqrt{2}}, \tag{3}$$

where $l(i,j)$ is the distance from the entry (i,j) to the diagonal line and the standard deviation parameter $\sigma = 1$. The values of $\mathbf{T}$ peak on the diagonal and gradually decrease along the direction perpendicular to the diagonal.

In this work, we adopt the above temporal order-preserving prior for few-shot video classification. Let $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{T}}$ denote the row-wise normalized version of $\mathbf{D}$ and $\mathbf{T}$ respectively, as:

$$\tilde{\mathbf{D}}(i,j) = \frac{\exp^{\mathbf{D}(i,j)}}{\sum_{k=1}^{M} \exp^{\mathbf{D}(i,k)}}, \quad \tilde{\mathbf{T}}(i,j) = \frac{\mathbf{T}(i,j)}{\sum_{k=1}^{M} \mathbf{T}(i,k)}. \tag{4}$$

We define the temporal similarity score between $\mathbf{X}$ and $\mathbf{Y}$ as the negative Kullback-Leibler (KL) divergence between $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{T}}$:

$$\text{sim}_t(f_\theta(\mathbf{X}), f_\theta(\mathbf{Y})) = -KL(\tilde{\mathbf{D}}||\tilde{\mathbf{T}}) = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{M} \tilde{\mathbf{D}}(i,j) \log \frac{\tilde{\mathbf{D}}(i,j)}{\tilde{\mathbf{T}}(i,j)}. \tag{5}$$

Intuitively, $\text{sim}_t(f_\theta(\mathbf{X}), f_\theta(\mathbf{Y}))$ encourages the temporal alignment between $\mathbf{X}$ and $\mathbf{Y}$ to be as similar as possible to the temporal order-preserving alignment $\mathbf{T}$. Fig. 2(b) summarizes the steps to compute our temporal similarity score.

3.3 Training and Testing

Training. We employ the global training with prototype-based classifiers [38] in our approach. A set of prototypes $\mathbf{W} = [\mathbf{W}_1, \ldots, \mathbf{W}_{N_b}]$ are initialized randomly, where each prototype $\mathbf{W}_p = [\mathbf{w}_p^1, \ldots, \mathbf{w}_p^M] \in \mathbb{R}^{C \times M}$ represents a class in the base set. Note that in contrast to few-shot image classification [38], where each prototype is a single feature vector, in our approach for few-shot video classification, each prototype is a sequence of M feature vectors. To learn $\mathbf{W}$, we adopt the below supervised loss:

$$\mathcal{L}_{Sup} = -\mathbb{E}_{(\mathbf{X},y)\sim\mathcal{D}_b} \log \frac{\exp^{\text{sim}_a(f_\theta(\mathbf{X}),\mathbf{W}_y))}}{\sum_{p=1}^{N_b} \exp^{\text{sim}_a(f_\theta(\mathbf{X}),\mathbf{W}_p))}} - \alpha\mathbb{E}_{(\mathbf{X},y)\sim\mathcal{D}_b}\text{sim}_t(f_\theta(\mathbf{X}), \mathbf{W}_y). \tag{6}$$

Here, α is the balancing weight between the two terms, and the appearance and temporal similarity scores are computed between the features of a video and the prototype of a class. We empirically observe that the temporal order-preserving

prior is effective for Something-Something V2 but not much for Kinetics. This is likely because the actions in Something-Something V2 are order-sensitive, whereas those in Kinetics are not. Therefore, we use both terms in Eq. 6 for training on Something-Something V2 (i.e., $\alpha = 0.05$) but only the first term for training on Kinetics (i.e., $\alpha = 0$).

Solely training with the above supervised loss can lead to a trivial solution that the model only learns discriminative features for each class (i.e., the M feature vectors within each prototype are similar). To avoid such cases, we add another loss which minimizes the entropy of $\tilde{\mathbf{D}}$:

$$\mathcal{L}_{Info} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{M}\tilde{\mathbf{D}}(i,j)\log\tilde{\mathbf{D}}(i,j). \tag{7}$$

The final training loss is a combination of the above losses and is written as:

$$\mathcal{L} = \mathcal{L}_{Sup} + \nu\mathcal{L}_{Info}. \tag{8}$$

Here, $\nu = 0.1$ is the balancing weight.

Testing. At test time, we discard the global prototype-based classifiers, and keep the feature extractor. Given an N-way K-shot episode, we first extract the features of all support and query samples. We then initialize N prototypes by the average features of support samples from the corresponding classes, i.e., $\mathbf{W}_c = \frac{1}{K}\sum_{\mathbf{X}\in\mathcal{S}_c} f_\theta(\mathbf{X})$, where $\mathcal{S}_c$ is the support set of the c-th class. Given the prototypes, we define the predictive distribution over classes of each video sample $p(c|\mathbf{X},\mathbf{W}), c \in \{1,\dots,N\}$, as:

$$p(c|\mathbf{X},\mathbf{W}) = (1-\beta)\frac{\exp^{\text{sim}_a(f_\theta(\mathbf{X}),\mathbf{W}_c)}}{\sum_{j=1}^{N}\exp^{\text{sim}_a(f_\theta(\mathbf{X}),\mathbf{W}_j)}} + \beta\frac{\exp^{\text{sim}_t(f_\theta(\mathbf{X}),\mathbf{W}_c)}}{\sum_{j=1}^{N}\exp^{\text{sim}_t(f_\theta(\mathbf{X}),\mathbf{W}_j)}}, \tag{9}$$

where β is the balancing weight between the two terms. The above predictive distribution is a combination of the *softmax* predictions based on appearance and temporal similarity scores. As we discussed previously, since the actions on Something-Something V2 are order-sensitive but those on Kinetics are not, the temporal order-preserving is effective for Something-Something V2 but not for Kinetics. Therefore, we set $\beta = 0.5$ for Something-Something V2 and $\beta = 0$ for Kinetics. As we empirically show in Sec. 4.1, these settings yield good results.

3.4 Prototype Refinement

The prototypes can be further refined with support samples (in the inductive setting) or with both support and query samples (in the transductive setting) before being used for classification.

Inductive Setting. For the inductive inference, we finetune the prototypes on the support set with the following cross-entropy loss:

$$\mathcal{L}_{inductive} = -\mathbb{E}_{(\mathbf{X},y)\sim\mathcal{D}^s} \log \frac{\exp^{\text{sim}_a(f_\theta(\mathbf{X}),\mathbf{W}_y)}}{\sum_{i=1}^{N} \exp^{\text{sim}_a(f_\theta(\mathbf{X}),\mathbf{W}_i)}}. \tag{10}$$

Transductive Setting. We introduce a transductive inference step that utilizes unsupervised information from the query set to finetune the prototypes. The transductive inference typically has a form of Soft K-means [33] with a novel assignment function. For each update iteration, we refine the prototypes with the weighted sums of the support and query samples:

$$\mathbf{W}_c = \frac{\sum_{\mathbf{X}\sim\mathcal{S}} f_\theta(\mathbf{X}) z(\mathbf{X},c) + \sum_{\mathbf{X}\sim\mathcal{Q}} f_\theta(\mathbf{X}) z(\mathbf{X},c)}{\sum_{\mathbf{X}\sim\mathcal{S}} z(\mathbf{X},c) + \sum_{\mathbf{X}\sim\mathcal{Q}} z(\mathbf{X},c)}, \tag{11}$$

where $\mathcal{S}$ and $\mathcal{Q}$ are the support and query sets, respectively. The assignment function $z(\mathbf{X},c)$ simply returns the c-th element of the one-hot label vector of $\mathbf{X}$ if the sample is from the support set. For query samples, we directly use the predictive distribution in Eq. 9, i.e., $z(\mathbf{X},c) = p(c|\mathbf{X},\mathbf{W})$ for $\mathbf{X} \in \mathcal{Q}$. In our experiments, the prototypes are updated for 10 iterations.

4 Experiments

There are two parts of experiments in this section. In the first part (Sec. 4.1), we conduct ablation studies to show the effectiveness of appearance and temporal similarity scores. In the second part (Sec. 4.2), we compare our proposed method to state-of-the art methods on two widely used datasets: Kinetics [20], and Something-Something V2 [15]. The evaluations are conducted on both inductive and transductive settings.

Datasets. We perform experiments on two standard few-shot video classification datasets: the few-shot versions of Kinetics [20] and Something-Something V2 [15]. Kinetics [20] contains 10,000 videos, while Something-Something V2 [15] has 71,796 videos. These two datasets are split into 64 training classes, 16 validation classes, and 20 testing classes, following the splits from [56] and [5].

Implementation Details. For fair comparisons, we follow the preprocessing steps from prior works [5,57,54]. In particular, we first resize the frames of a particular video to 256×256 and perform random cropping (for training) or central cropping (for testing) a 224×224 region from the frames. The number of sampled segments or frames for each video is $M = 8$. We use the same network architecture as previous works, which is a ResNet-50 pretrained on ImageNet [35]. Stochastic gradient descent (SGD) with a momentum of 0.9 is adopted as our optimizer. The model is trained for 25 epochs with an initial learning rate of 0.001 and a weight decay of 0.1 at epoch 20. In the inference stage, we report mean accuracy with 95% confidence interval on 10,000 random episodes.

Table 1. Comparison to different aggregation methods in the **inductive** setting on the Kinetics and Something-Something V2 datasets.

Method	Kinetics		Something V2	
	1-shot	5-shot	1-shot	5-shot
OT	61.71 ± 0.41	77.16 ± 0.37	35.60 ± 0.41	47.82 ± 0.44
Max	73.46 ± 0.38	$\mathbf{87.67 \pm 0.29}$	40.97 ± 0.41	58.77 ± 0.43
Ours	$\mathbf{74.26 \pm 0.38}$	87.40 ± 0.30	$\mathbf{43.82 \pm 0.42}$	$\mathbf{61.07 \pm 0.42}$

4.1 Ablation Study

The Effectiveness of Appearance and Temporal Similarity Scores. In this experiment, we compare our appearance and temporal similarity scores with several methods for aggregating the appearance similarity matrix $\mathbf{D}$ in Eq. 2, namely the optimal transport (OT) and the max operator. For OT, we use the negative appearance similarity matrix $-\mathbf{D}$ as the cost matrix and employ the equal partition constraint. It optimally matches frames of two videos without considering temporal information. For the max operator, we replace the log-sum-exp operator in Eq. 2 with the max operator. Tab. 1 shows their results on Kinetics and Something-Something V2 datasets. Our method outperforms the other two similarity functions by large margins on both datasets with the exception that the max variant performs slightly better than our method in the Kinetics 5-way 5-shot setting (87.67% as compared to 87.40% of our method).

Table 2. Comparison of different training losses in the **inductive** and **transductive** settings on the Kinetics and Something-Something V2 datasets.

$\mathcal{L}_{Sup}$	$\mathcal{L}_{Info}$	Transd.	Kinetics		Something V2	
			1-shot	5-shot	1-shot	5-shot
✓			74.25 ± 0.38	87.15 ± 0.30	41.84 ± 0.42	57.26 ± 0.43
✓	✓		74.26 ± 0.38	87.40 ± 0.30	43.82 ± 0.42	61.07 ± 0.42
✓		✓	82.71 ± 0.44	92.91 ± 0.31	47.02 ± 0.54	67.56 ± 0.56
✓	✓	✓	$\mathbf{83.08 \pm 0.44}$	$\mathbf{93.33 \pm 0.30}$	$\mathbf{48.67 \pm 0.54}$	$\mathbf{69.42 \pm 0.55}$

The Effectiveness of Training Losses. We perform analysis of our training losses in Sec. 3.3 on both Kinetics and Something-Something V2 datasets. Results are shown in Tab. 2. As mentioned earlier, adding $\mathcal{L}_{Info}$ in the training phase helps avoiding trivial solutions and hence consistently improves the performance on both Kinetics and Something-Something V2 datasets. Using the complete training loss with the transductive setting yields the best performance on both Kinetics and Something-Something V2 datasets.

The Relative Importance of Appearance and Temporal Cues. In this experiment, we investigate the effectiveness of the hyperparameter β in the inductive and transductive settings. More specifically, we train our model with the loss in Eq. 6 and evaluate the result on the validation set with different values of β. The results of the inductive and transductive settings are shown in Tab. 3.

Table 3. Ablation study on the relative importance of the appearance and temporal terms in computing the predictive distribution and the assignment function of the **inductive** and **transductive** inference on the Kinetics and Something-Something V2 datasets.

		Kinetics		Something V2	
β	**Transd.**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
1.00		21.09 ± 0.34	27.00 ± 0.39	25.28 ± 0.37	33.57 ± 0.41
0.90		24.02 ± 0.36	41.75 ± 0.44	28.09 ± 0.38	43.21 ± 0.43
0.80		28.94 ± 0.38	62.66 ± 0.43	32.24 ± 0.40	54.45 ± 0.43
0.70		39.75 ± 0.42	79.99 ± 0.35	39.13 ± 0.42	62.73 ± 0.42
0.60		58.87 ± 0.43	86.37 ± 0.30	45.40 ± 0.43	65.99 ± 0.41
0.50	✗	74.05 ± 0.37	87.33 ± 0.29	48.05 ± 0.43	$\mathbf{66.91 \pm 0.40}$
0.40		75.16 ± 0.37	$\mathbf{87.38 \pm 0.29}$	$\mathbf{48.55 \pm 0.42}$	66.83 ± 0.40
0.30		75.34 ± 0.36	87.38 ± 0.29	48.54 ± 0.42	66.64 ± 0.41
0.20		75.37 ± 0.36	87.37 ± 0.29	48.36 ± 0.43	66.32 ± 0.41
0.10		$\mathbf{75.38 \pm 0.36}$	87.36 ± 0.29	48.21 ± 0.43	66.06 ± 0.41
0.00		75.37 ± 0.36	87.35 ± 0.29	48.13 ± 0.43	65.77 ± 0.41
1.00		54.83 ± 0.49	66.61 ± 0.50	39.53 ± 0.48	51.46 ± 0.51
0.90		63.88 ± 0.49	80.48 ± 0.43	43.00 ± 0.50	58.45 ± 0.53
0.80		72.38 ± 0.48	88.71 ± 0.36	46.56 ± 0.53	65.33 ± 0.54
0.70		78.09 ± 0.45	91.49 ± 0.32	50.03 ± 0.54	70.46 ± 0.53
0.60	✓	80.73 ± 0.44	92.66 ± 0.31	52.36 ± 0.55	73.42 ± 0.52
0.50		82.34 ± 0.43	93.24 ± 0.30	53.65 ± 0.55	74.77 ± 0.52
0.40		83.29 ± 0.43	93.51 ± 0.29	54.46 ± 0.55	75.42 ± 0.52
0.30		83.81 ± 0.43	93.59 ± 0.29	$\mathbf{54.67 \pm 0.55}$	$\mathbf{75.47 \pm 0.52}$
0.20		84.06 ± 0.43	$\mathbf{93.60 \pm 0.29}$	54.75 ± 0.55	75.25 ± 0.52
0.10		84.33 ± 0.42	93.55 ± 0.29	54.66 ± 0.55	74.77 ± 0.52
0.00		$\mathbf{84.36 \pm 0.41}$	93.42 ± 0.29	54.4 ± 0.54	73.97 ± 0.51

The two table shows that for order-sensitive actions in Something-Something V2, balancing between the importance of appearance and temporal scores gives the best performance. In particular, for the inductive setting, our approach optimally achieves 48.55% for 1-shot with $\beta = 0.4$, whereas the best result for 5-shot is 66.91% with $\beta = 0.5$. Next, $\beta = 0.3$ gives the best results for both 1-shot and 5-shot in the transductive setting, achieving 54.76% and 75.47% respectively. Jointly considering appearance and temporal cues consistently improves the model performance as compared to the appearance-only version ($\beta = 0.0$).

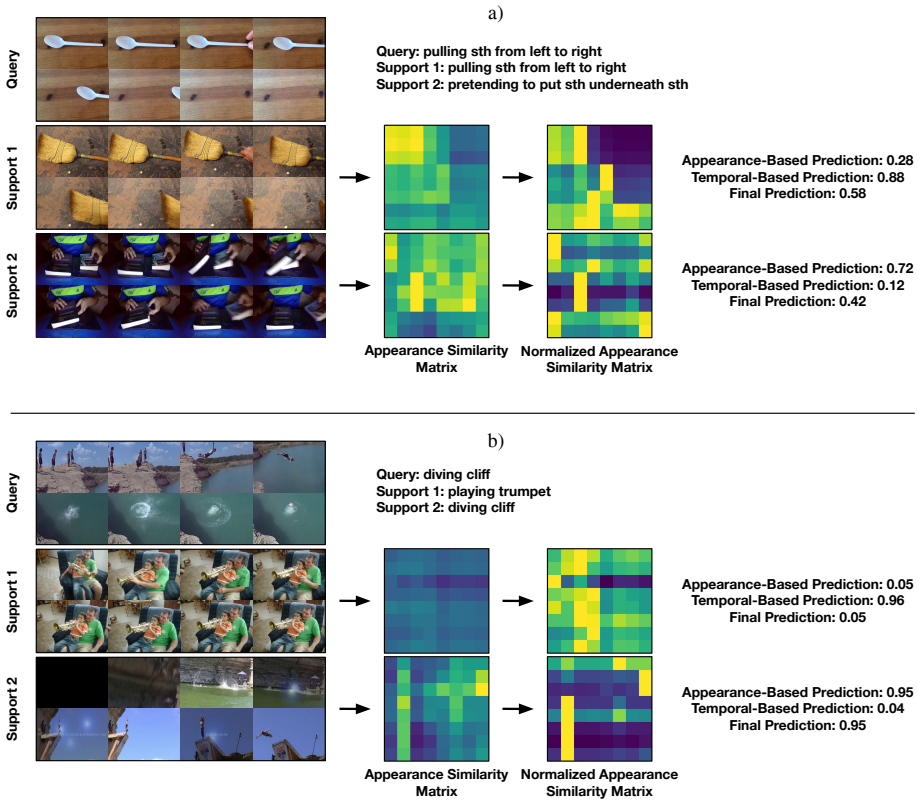

Fig. 3. Qualitative Results of 2-way 1-shot Tasks on Something-Something V2 and Kinetics. For each task, we present the appearance similarity matrix $\mathbf{D}$ between the query video and each support video in the second column. In the third column, we show the row-wise normalized version $\tilde{\mathbf{D}}$. Finally, we show the predictions of the two similarity scores and the final prediction. Ground truth class labels are shown at the top. a) Results from Something-Something V2. b) Results from Kinetics.

In contrast, temporal information does not show much benefits for an order-insensitive dataset like Kinetics. For small values of β, there are minor changes in the performance of our approach for both inductive and transductive settings on Kinetics. In the worst case, the temporal-only version ($\beta = 1.0$) produces nearly random guesses of 21.09% in the 1-shot inductive setting.

We show some qualitative results in Fig. 3. We respectively perform inductive inferences on 2-way 1-shot tasks of Something-Something V2 and Kinetics datasets with $\beta = 0.5$ and $\beta = 0.0$ respectively. On Something-Something V2 (Fig. 3(a)), we observe that appearance or temporal cues can misclassify query samples sometimes, but utilizing both appearance or temporal cues gives correct classifications. On the other hand, the actions on Kinetics ((Fig. 3(b)) are not

order-sensitive, and hence the temporal similarity score is not meaningful, which agrees with the results in Tab. 3.

Table 4. Comparison to the state-of-the-art methods in the **inductive** setting on the Kinetics and Something-Something V2 datasets. Results of Meta-Baseline [38] and CMN [56] are reported from [57]. † denotes results from our re-implementation.

Method	Kinetics		Something V2	
	1-shot	5-shot	1-shot	5-shot
Meta-Baseline [38]	64.03 ± 0.41	80.43 ± 0.35	37.31 ± 0.41	48.28 ± 0.44
CMN [56]	65.90 ± 0.42	82.72 ± 0.34	40.62 ± 0.42	51.90 ± 0.44
OTAM [5]	$73.00 \pm n/a$	$85.80 \pm n/a$	$42.80 \pm n/a$	$52.30 \pm n/a$
Baseline Plus [57]†	70.48 ± 0.40	82.67 ± 0.33	43.05 ± 0.41	57.50 ± 0.43
ITANet [54]	73.60 ± 0.20	84.30 ± 0.30	$\mathbf{49.20 \pm 0.20}$	$\mathbf{62.30 \pm 0.30}$
Ours	$\mathbf{74.26 \pm 0.38}$	$\mathbf{87.40 \pm 0.30}$	43.82 ± 0.42	61.07 ± 0.42

4.2 Comparison with Previous Methods

We compare our approach against previous methods [38,56,33,26,5,54,57] in both inductive and transductive settings. More specifically, in the inductive setting, we first consider Prototypical Network [38] from few-shot image classification, which is re-implemented by [57]. In addition, CMN [56], OTAM [5], Baseline Plus [57], and ITANet [54], which are previous methods designed to tackle few-shot video classification, are also considered. The results of Prototypical Network (namely, Meta-Baseline) and CMN are taken from [57]. We re-implement Baseline Plus [57]. The results of OTAM and ITANet are taken from the original papers. Competing methods in the transductive setting include clustering-based methods from few-shot image classification, namely Soft K-means [33], Bayes K-means [26], and Mean-shift [26].

Inductive. We first consider the inductive setting. The results are presented in Tab. 4. As can be seen from the table, our method achieves the best results on the Kinetics dataset. It outperforms all the competing methods by around 1% for 1-shot setting and over 3% for 5-shot setting, establishing the new state of the art. In addition, our method performs comparably with previous works on Something-Something V2, outperforming all the competing methods except for ITANet, which further adopts additional layers on top of the ResNet-50 for self-attention modules.

Transductive. Next, we consider the transductive setting (shown in Tab. 5). As can be seen in Tab. 5, our method outperforms all the competing methods by large margins, which are around 9% for both 1-shot and 5-shot settings on Kinetics and 2% and 5% for 1-shot and 5-shot settings respectively on Something-Something V2.

Table 5. Comparison to the state-of-the-art methods in **transductive** setting on the Kinetics and Something-Something V2 datasets. Results of other methods are from our re-implementation on the trained feature extractor of [57].

Method	Kinetics		Something V2	
	1-shot	5-shot	1-shot	5-shot
Soft K-means [33]	74.21 ± 0.40	84.13 ± 0.33	46.46 ± 0.46	64.93 ± 0.45
Bayes K-means [26]	70.66 ± 0.40	81.21 ± 0.34	43.15 ± 0.41	59.48 ± 0.42
Mean-shift [26]	70.52 ± 0.40	82.31 ± 0.34	43.15 ± 0.41	60.03 ± 0.43
Ours	$\mathbf{83.08 \pm 0.44}$	$\mathbf{93.33 \pm 0.30}$	$\mathbf{48.67 \pm 0.54}$	$\mathbf{69.42 \pm 0.55}$

5 Limitation Discussion

We propose two similarity functions for aligning appearance and temporal cues of videos. While our results are promising in both inductive and transductive experiments, there remain some limitations. Firstly, our temporal order-preserving prior does not work for all datasets. Utilizing a permutation-aware temporal prior [27] would be an interesting next step. Secondly, we have not leveraged spatial information in our approach yet. Such spatial information could be important for scenarios like modeling left-right concepts. We leave this investigation as our future work.

6 Conclusion

We propose, in this paper, a novel approach for few-shot video classification via appearance and temporal alignments. Specifically, our approach performs frame-level feature alignment to compute the appearance similarity score between the query and support videos, while utilizing temporal order-preserving priors to calculate the temporal similarity score between the videos. The proposed similarity scores are then used across different stages of our few-shot video classification framework, namely prototype-based training and testing, and inductive and transductive prototype enhancement. We show that our similarity scores are most effective on temporal order-sensitive datasets such as Something-Something V2, while our approach produces comparable or better results than previous few-shot video classification methods on both Kinetics and Something-Something V2 datasets. To the best of our knowledge, our work is the first to explore transductive few-shot video classification, which could facilitate more future works in this direction.

References

1. Ben-Ari, R., Nacson, M.S., Azulai, O., Barzelay, U., Rotman, D.: Taen: Temporal aware embedding network for few-shot action recognition. In: CVPR (2021) 2, 4
2. Bo, Y., Lu, Y., He, W.: Few-shot learning of video action recognition only based on video contents. In: WACV (2020) 2, 4
3. Boudiaf, M., Ziko, I., Rony, J., Dolz, J., Piantanida, P., Ben Ayed, I.: Information maximization for few-shot learning. NeurIPS (2020) 4
4. Cao, C., Li, Y., Lv, Q., Wang, P., Zhang, Y.: Few-shot action recognition with implicit temporal alignment and pair similarity optimization. CVIU (2021) 2, 4
5. Cao, K., Ji, J., Cao, Z., Chang, C.Y., Niebles, J.C.: Few-shot video classification via temporal alignment. In: CVPR (2020) 2, 3, 4, 5, 6, 9, 13
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017) 1
7. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. NeurIPS (2013) 2
8. Cuturi, M., Blondel, M.: Soft-dtw: a differentiable loss function for time-series. In: ICML (2017) 2
9. Das, R., Wang, Y.X., Moura, J.M.: On the importance of distractors for few-shot classification. In: ICCV (2021) 2, 4
10. Fei, N., Gao, Y., Lu, Z., Xiang, T.: Z-score normalization, hubness, and few-shot learning. In: ICCV (2021) 2
11. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. arXiv preprint arXiv:1703.03400 (2017) 1, 3
12. Fu, Y., Zhang, L., Wang, J., Fu, Y., Jiang, Y.G.: Depth guided adaptive meta-fusion network for few-shot video recognition. In: ACM MM (2020) 2, 4
13. Geetha, P., Narayanan, V.: A survey of content-based video retrieval (2008) 1
14. Ghaffari, S., Saleh, E., Forsyth, D., Wang, Y.X.: On the importance of firth bias reduction in few-shot classification. In: ICLR (2022) 2
15. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: ICCV (2017) 3, 9
16. Haresh, S., Kumar, S., Coskun, H., Syed, S.N., Konin, A., Zia, Z., Tran, Q.H.: Learning by aligning videos in time. In: CVPR (2021) 3, 6
17. Haresh, S., Kumar, S., Zia, M.Z., Tran, Q.H.: Towards anomaly detection in dashcam videos. In: 2020 IEEE Intelligent Vehicles Symposium (IV). pp. 1407–1414. IEEE 1
18. Hou, R., Chang, H., Ma, B., Shan, S., Chen, X.: Cross attention network for few-shot classification. arXiv preprint arXiv:1910.07677 (2019) 3, 4
19. Kang, D., Kwon, H., Min, J., Cho, M.: Relational embedding for few-shot classification. In: ICCV (2021) 3
20. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017) 3, 9
21. Khan, H., Haresh, S., Ahmed, A., Siddiqui, S., Konin, A., Zia, M.Z., Tran, Q.H.: Timestamp-supervised action segmentation with graph convolutional networks. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2022) 1

22. Konin, A., Syed, S.N., Siddiqui, S., Kumar, S., Tran, Q.H., Zia, M.Z.: Retroactivity: Rapidly deployable live task guidance experiences. In: IEEE International Symposium on Mixed and Augmented Reality Demonstration (2020) 1
23. Kumar, S., Haresh, S., Ahmed, A., Konin, A., Zia, M.Z., Tran, Q.H.: Unsupervised activity segmentation by joint representation learning and online clustering. In: CVPR (2022) 3, 6
24. Le, D., Nguyen, K.D., Nguyen, K., Tran, Q.H., Nguyen, R., Hua, B.S.: Poodle: Improving few-shot learning via penalizing out-of-distribution samples. NeurIPS (2021) 2, 4
25. Li, Z., Zhou, F., Chen, F., Li, H.: Meta-sgd: Learning to learn quickly for few-shot learning. arXiv preprint arXiv:1707.09835 (2017) 1, 3
26. Lichtenstein, M., Sattigeri, P., Feris, R., Giryes, R., Karlinsky, L.: Tafssl: Task-adaptive feature sub-space learning for few-shot classification. In: ECCV (2020) 13, 14
27. Liu, W., Tekin, B., Coskun, H., Vineet, V., Fua, P., Pollefeys, M.: Learning to align sequential actions in the wild. arXiv preprint arXiv:2111.09301 (2021) 14
28. Lu, S., Ye, H.J., Zhan, D.C.: Few-shot action recognition with compromised metric via optimal transport. arXiv preprint arXiv:2104.03737 (2021) 2, 4, 5
29. Ma, C., Huang, Z., Gao, M., Xu, J.: Few-shot learning via dirichlet tessellation ensemble. In: ICLR (2022) 2
30. Munkhdalai, T., Sordoni, A., Wang, T., Trischler, A.: Metalearned neural memory. NeurIPS (2019) 3
31. Oreshkin, B., Rodríguez López, P., Lacoste, A.: Tadam: Task dependent adaptive metric for improved few-shot learning. NeurIPS (2018) 1, 3
32. Qiao, S., Liu, C., Shen, W., Yuille, A.L.: Few-shot image recognition by predicting parameters from activations. In: CVPR (2018) 1, 3
33. Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J.B., Larochelle, H., Zemel, R.S.: Meta-learning for semi-supervised few-shot classification. arXiv preprint arXiv:1803.00676 (2018) 4, 9, 13, 14
34. Rodriguez, M., Sivic, J., Laptev, I., Audibert, J.Y.: Data-driven crowd analysis in videos. In: 2011 International Conference on Computer Vision. pp. 1235–1242. IEEE (2011) 1
35. 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) 9
36. Rusu, A.A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., Hadsell, R.: Meta-learning with latent embedding optimization. arXiv preprint arXiv:1807.05960 (2018) 1, 3
37. Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T.: Meta-learning with memory-augmented neural networks. In: ICML (2016) 1, 3
38. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: NeurIPS (2017) 1, 3, 7, 13
39. Snoek, C.G., Worring, M.: Concept-based video retrieval. Now Publishers Inc (2009) 1
40. Su, B., Hua, G.: Order-preserving wasserstein distance for sequence matching. In: CVPR (2017) 3, 6
41. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6479–6488 (2018) 1
42. 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) 1, 3

43. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV (2015) 1
44. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: CVPR (2018) 1
45. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: NeurIPS (2016) 1, 3
46. Wang, R., Pontil, M., Ciliberto, C.: The role of global labels in few-shot classification and how to infer them. In: NeurIPS (2021) 2
47. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018) 1
48. Wertheimer, D., Tang, L., Hariharan, B.: Few-shot classification with feature map reconstruction networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8012–8021 (2021) 4
49. Wu, J., Zhang, T., Zhang, Y., Wu, F.: Task-aware part mining network for few-shot learning. In: ICCV (2021) 3
50. Yang, S., Liu, L., Xu, M.: Free lunch for few-shot learning: Distribution calibration. arXiv preprint arXiv:2101.06395 (2021) 2, 4
51. Zhan, B., Monekosso, D.N., Remagnino, P., Velastin, S.A., Xu, L.Q.: Crowd analysis: a survey. Machine Vision and Applications **19**(5), 345–357 (2008) 1
52. Zhang, C., Cai, Y., Lin, G., Shen, C.: Deepemd: Differentiable earth mover's distance for few-shot learning. arXiv preprint arXiv:2003.06777 (2020) 3
53. Zhang, H., Zhang, L., Qi, X., Li, H., Torr, P.H., Koniusz, P.: Few-shot action recognition with permutation-invariant attention. In: ECCV (2020) 2, 4
54. Zhang, S., Zhou, J., He, X.: Learning implicit temporal alignment for few-shot video classification. arXiv preprint arXiv:2105.04823 (2021) 9, 13
55. Zhang, X., Meng, D., Gouk, H., Hospedales, T.M.: Shallow bayesian meta learning for real-world few-shot recognition. In: ICCV (2021) 3
56. Zhu, L., Yang, Y.: Compound memory networks for few-shot video classification. In: ECCV (2018) 4, 9, 13
57. Zhu, Z., Wang, L., Guo, S., Wu, G.: A closer look at few-shot video classification: A new baseline and benchmark. arXiv preprint arXiv:2110.12358 (2021) 2, 5, 9, 13, 14
58. Ziko, I., Dolz, J., Granger, E., Ayed, I.B.: Laplacian regularized few-shot learning. In: ICML (2020) 4