Spatio-temporal Relation Modeling for Few-shot Action Recognition

Anirudh Thatipelli[1] Sanath Narayan[2] Salman Khan[1,4]
Rao Muhammad Anwer[1,3] Fahad Shahbaz Khan[1,5] Bernard Ghanem[6]
[1]Mohamed Bin Zayed University of Artificial Intelligence [2]Inception Institute of Artificial Intelligence [3]Aalto University
[4]Australian National University [5]CVL, Linköping University [6]King Abdullah University of Science & Technology

Abstract

We propose a novel few-shot action recognition framework, STRM, which enhances class-specific feature discriminability while simultaneously learning higher-order temporal representations. The focus of our approach is a novel spatio-temporal enrichment module that aggregates spatial and temporal contexts with dedicated local patch-level and global frame-level feature enrichment sub-modules. Local patch-level enrichment captures the appearance-based characteristics of actions. On the other hand, global frame-level enrichment explicitly encodes the broad temporal context, thereby capturing the relevant object features over time. The resulting spatio-temporally enriched representations are then utilized to learn the relational matching between query and support action sub-sequences. We further introduce a query-class similarity classifier on the patch-level enriched features to enhance class-specific feature discriminability by reinforcing the feature learning at different stages in the proposed framework. Experiments are performed on four few-shot action recognition benchmarks: Kinetics, SSv2, HMDB51 and UCF101. Our extensive ablation study reveals the benefits of the proposed contributions. Furthermore, our approach sets a new state-of-the-art on all four benchmarks. On the challenging SSv2 benchmark, our approach achieves an absolute gain of 3.5% *in classification accuracy, as compared to the best existing method in the literature. Our code and models are available at* https://github.com/Anirudh257/strm.

1. Introduction

Few-shot (FS) action recognition is a challenging computer vision problem, where the task is to classify an unlabelled query video into one of the action categories in the support set having limited samples per action class. The problem setting is particularly relevant for fine-grained action recognition [11], since it is challenging to collect sufficient labelled examples [4, 5]. Most existing FS action recognition methods typically search for either a single support video [31] or an average representation of a support class [2, 3]. However, these approaches utilize only frame-level representations and do not explicitly exploit video sub-sequences for temporal relationship modeling.

In the context of FS action recognition, modeling temporal relationships between a query video and limited support actions is a major challenge, since actions are typically performed at various speeds and occur at different time instants (temporal offsets). Further, video representations are desired to encode the relevant information from multiple sub-actions that constitute an action for enhanced matching between query and support videos. Moreover, an effective representation of spatial and temporal contexts of actions is crucial to distinguish fine-grained classes requiring temporal relational reasoning, where actions can be performed with different objects in various backgrounds, *e.g.*, *spilling something behind something*.

The aforementioned problem of temporal relationship modeling is recently explored by Temporal-Relational CrossTransformers (TRX) [19], which compares the sub-sequences of query and support videos in a part-based manner to tackle the issue of varying speed and offsets of actions. Additionally, TRX models complex higher-order temporal relations by representing sub-sequences as tuples with different cardinalities. However, TRX struggles in the case of actions performed with different objects and background (see Fig. 1). This is likely due to not explicitly utilizing the available rich spatio-temporal contextual information during temporal relationship modeling. Furthermore, the tuple representations in TRX are fixed requiring a separate CrossTransformer [7] branch per cardinality, which affects the model flexibility. Here, we set out to collectively address the above issues while modeling temporal relationships between query and limited support actions.

In this work, we argue that both local patch features in a frame and global frame features in a video are desirable cues to effectively enrich the encoding of spatial as well as temporal contextual information. Such feature enrichment improves class-specific discriminability, enabling focus on relevant objects and their corresponding motion in a video.

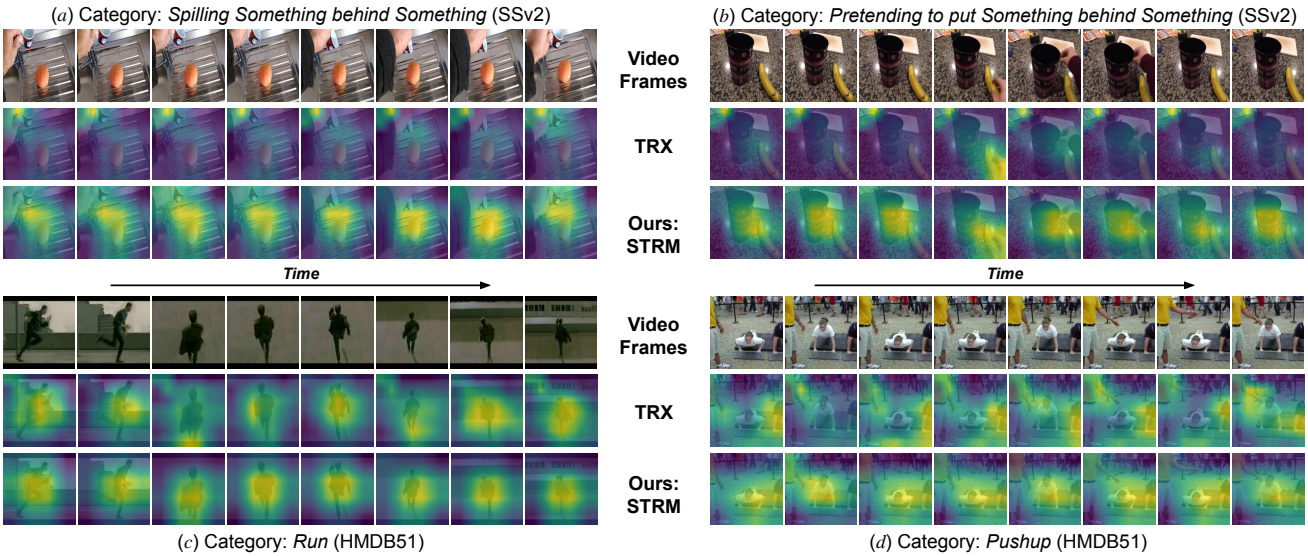


Figure 1. **Example attention map visualizations** obtained from the recently introduced TRX [19] and our proposed STRM approach on four examples from the SSV2 and HMDB51 test set. The attention maps measure the activation magnitude of latent features. TRX struggles in the case of spatial and temporal context variations that are commonly encountered in actions performed with different objects and backgrounds, *e.g.*, 5^{th} and 6^{th} frame from the left in (*b*), where the regions corresponding to actions are not emphasized. Similarly, while background region is also emphasized in 3^{rd} and 6^{th} frame from the left in (c), the action in the 2^{nd} and 3^{rd} frame from the left in (*d*) is not accurately captured due to the distractor motion from the moving hand of another person. Our STRM approach explicitly enhances class-specific feature discriminability through spatio-temporal context aggregation and intermediate latent feature classification. This leads to better matching between query and limited support action instances. Additional examples are presented in Fig. 5 and supplementary.

In addition, learning to classify feature representations at different stages is expected to reinforce the model to look for class-separable features, thereby further improving the class-specific discriminability. Moreover, this class-specific discriminability is attainable through a reduced set of cardinalities generated by the automatic learning of higher-order temporal relationships.

Contributions: We introduce an FS action recognition framework that comprises spatio-temporal enrichment and temporal relationship modeling modules along with a query-class similarity classifier. The spatio-temporal enrichment module comprises local patch-level enrichment (PLE) and global frame-level enrichment (FLE) sub-modules. The PLE enriches local patch features with spatial context by attending to all patches in a frame, in a sample-dependent manner, in order to capture the appearance-based similarities as well as dissimilarities among the action categories. On the other hand, the FLE enriches global frame features with temporal context by persistent relationship memory-based (sample-agnostic) aggregation that encompasses the entire receptive field in order to capture the relevant object motion in a video. The resulting enriched frame-level global representations are then utilized in the temporal relationship modeling (TRM) module to learn the temporal relations between query and support actions. Our TRM module does not rely on multiple cardinalities to model higher-order relations. Instead, it utilizes the spatio-temporal enrichment module to learn higher-order temporal representations at lower cardinalities. Moreover, we introduce a query-class similarity classifier that further enhances class-specific discriminability of the spatio-temporally enriched features by learning to classify representations from intermediate layer outputs.

We conduct extensive experiments on four FS action recognition benchmarks: Kinetics [4], SSv2 [11], HMDB51, [14] and UCF101 [23]. Our extensive ablations show that both the proposed spatio-temporal enrichment and query-class similarity classifier enhance feature discriminability, leading to significant improvements over the baseline. The spatio-temporal enrichment module further enables the modeling of temporal relationships using a single cardinality. Our approach outperforms existing FS action recognition methods in the literature on all four benchmarks. On the challenging SSv2 benchmark, our approach achieves classification accuracy of 68.1% with an absolute gain of 3.5% over the recently introduced TRX [19], when employing the ResNet-50 backbone. Fig. 1 shows a comparison of our approach with TRX, in terms of attention map visualizations, on examples from SSv2 and HMDB51.

2. Preliminaries

Problem Formulation: The goal of few-shot (FS) action recognition is to classify an unlabelled query video into one

of the C action classes in the 'support set' comprising K labelled instances for each class that is unseen during training. To this end, let $Q = \{q_1, \cdots, q_L\}$ denote a query video of L frames that is to be classified into a class $c \in C$. Moreover, let S^c be the support set of K videos for an action class c with the k^{th} video denoted as $S^c_k = \{s^c_{k1}, \cdots, s^c_{kL}\}$. For simplicity, we represent each video as a sequence of uniformly sampled L frames. In this work, we follow an episodic training paradigm as in [16], where few-shot tasks are randomly sampled from the training set for learning the C-way K-shot classification task in each episode. Next, we describe the baseline FS action recognition framework.

2.1. Baseline FS Action Recognition Framework

In this work, we adopt as baseline the recently introduced Temporal-relational CrossTransformer (TRX) [19] method, which has shown to achieve state-of-the-art performance on multiple action recognition benchmarks. The TRX classifies a query video by matching it with the actions occurring at different speeds and instants in the support class videos using CrossTransformers [7]. First, for each sub-sequence in the query video, a query-specific class prototype is computed via an aggregation of all possible sub-sequences in the support videos of an action class. The aggregation weights are based on the cross-attention values between the query sub-sequence and support class sub-sequences. Afterwards, the distances between the embeddings of the sub-sequences of a query video and their corresponding query-specific class prototypes are averaged to obtain the distance of the query to a class.

The TRX method introduces hand-crafted representations to capture the higher-order temporal relationships, where sub-sequences are represented by tuples of different cardinalities based on the number of frames used for encoding a sub-sequence. For instance, with $\mathbf{e}_i \in \mathbb{R}^D$ as the i^{th} frame representation, a sub-sequence between t_i and t_j can be represented as a pair $(\mathbf{e}_i, \mathbf{e}_j) \in \mathbb{R}^{2D}$, a triplet $(\mathbf{e}_i, \mathbf{e}_k, \mathbf{e}_j) \in \mathbb{R}^{3D}$, a quartet $(\mathbf{e}_i, \mathbf{e}_k, \mathbf{e}_l, \mathbf{e}_j) \in \mathbb{R}^{4D}$ and so on, such that $1 \leq i < k < l < j \leq L$. For a tuple $t = (t_1, \cdots, t_\omega)$ of cardinality $\omega \in \Omega$, let $\mathbf{q}_t \in \mathbb{R}^{D'}$ be a value embedding of query $Q_t = [\mathbf{e}_{t_1}; \cdots; \mathbf{e}_{t_\omega}] \in \mathbb{R}^{\omega D}$ and $\mathbf{p}^c_t \in \mathbb{R}^{D'}$ be the query-cardinality-specific class prototype, obtained by the attention-based aggregation of value embeddings of support tuples $S^c_{kt} \in \mathbb{R}^{\omega D}$. Then, the distance between a query video Q and support set $\mathbf{S}^c$ over multiple cardinalities is given by,

$$\mathbf{T}(Q, \mathbf{S}^c) = \sum_{\omega \in \Omega} \frac{1}{|\Pi_\omega|} \sum_{t \in \Pi_\omega} \|\mathbf{q}_t - \mathbf{p}^c_t\|, \qquad (1)$$

where $\Pi_\omega = \{(t_1, \cdots, t_\omega) \in \mathbb{N}^\omega : 1 \leq t_1 < \cdots < t_\omega \leq L\}$ is the set of all possible tuples for cardinality ω. The distance $\mathbf{T}(\cdot, \cdot)$ from a query video to its ground-truth class is minimized by employing a standard cross-entropy loss during training. For further details, we refer to [19].

Limitations: As discussed above, TRX performs temporal relationship modeling between the query and support action sub-sequences. However, this modeling struggles in the case of spatial context variation (appearance change of relevant objects in query and support videos) and associated variation in temporal context (aggregation of spatial context across frames). Such variations are typically encountered in case of fine-grained action categories (see Fig. 1). Furthermore, TRX jointly employs multiple CrossTransformers, one for each different cardinality, to model higher-order temporal relationships based on different hand-crafted temporal representations of sub-sequences. Consequently, this results in a less flexible model requiring dedicated branches for different cardinalities in addition to involving a manual model-search over different Ω combinations to find the optimal Ω^*. Next, we present our proposed approach that aims to collectively treat the aforementioned issues.

3. Proposed STRM Approach

Motivation: Here, we introduce our few-shot (FS) action recognition framework, STRM, which strives to enhance class-specific feature discriminability while simultaneously mitigating the flexibility issue.

Feature Discriminability: Distinct from TRX that focuses *solely* on temporal relationship modeling, our approach emphasizes the importance of aggregating spatial *and* temporal context to effectively enrich the video sub-sequence representations before modeling the temporal relations. The local representation followed by learning rich spatial and temporal relationships enables enhanced feature discriminability, leading to an effective utilization of the limited samples available for FS action recognition.

Model Flexibility: As discussed earlier, TRX employs hand-crafted higher-order temporal representations of different cardinalities, thereby requiring a search over multiple combinations. Instead, our approach learns to model higher-order relations at lower cardinalities with reduced inductive-bias, in turn improving the model flexibility.

To collectively address both the above issues, we introduce an enrichment mechanism that targets enhanced feature discriminability of individual frames at a local patch-level (spatial) as well as the video itself at a global frame-level (temporal) while simultaneously learning higher-order temporal representations for improved flexibility.

3.1. Overall Architecture

Fig. 2 illustrates our overall FS action recognition framework, STRM. The L video frames are passed through an image-feature extractor, which outputs D-dimensional frame features with a spatial resolution $P \times P$. The frame features are then spatially flattened to obtain $\mathbf{x}_i \in \mathbb{R}^{P^2 \times D}$, $i \in [1, L]$, which are then input to our novel spatio-temporal enrichment module comprising patch-level and frame-level

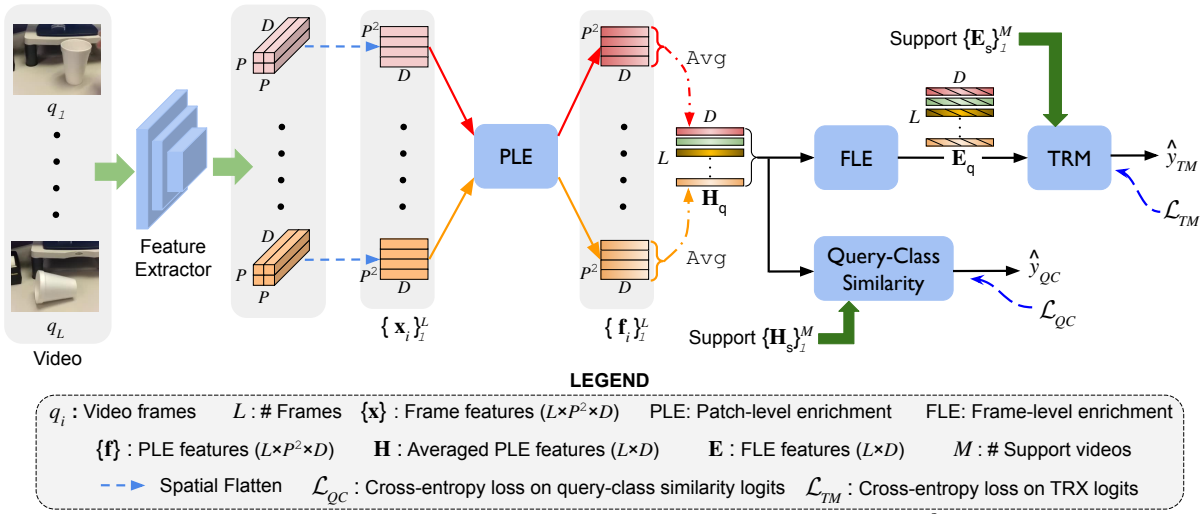


Figure 2. **Proposed STRM architecture** (Sec. 3.1). Spatially flattened D-dimensional features $\mathbf{x}_i \in \mathbb{R}^{P^2 \times D}$ are extracted for video frames q_i ($i \in [1, L]$). Here, P^2 is the number of patches. The features $\mathbf{x}_i$ are input to a patch-level enrichment (PLE, Sec. 3.2.1) block, which attends to the spatial context across patches in a frame and outputs spatially enriched features $\mathbf{f}_i \in \mathbb{R}^{P^2 \times D}$. Next, global representations $\mathbf{H} \in \mathbb{R}^{L \times D}$ are obtained by spatially averaging and temporally concatenating $\mathbf{f}_i$. These $\mathbf{H}$ are then input to a frame-level enrichment (FLE, Sec. 3.2.2) block, which models higher-order temporal representations by aggregating the temporal context of actions across frames in a video. The resulting *spatio-temporally enriched* features $\mathbf{E} \in \mathbb{R}^{L \times D}$ of query and support videos are then input to the TRM, which models the temporal relationships between them. Moreover, a query-class similarity classifier (Sec. 3.3) on the global representations $\mathbf{H}$ reinforces the network to learn class-discriminative features at different stages. Our framework is learned jointly using $\mathcal{L}_{TM}$ and $\mathcal{L}_{QC}$.

enrichment sub-modules to obtain class-discriminative representations. The patch-level enrichment (PLE) sub-module enhances the patch features locally by attending to the spatial context in each frame and outputs spatially enriched features $\mathbf{f}_i \in \mathbb{R}^{P^2 \times D}$ per frame. The $\mathbf{f}_i$'s are spatially averaged to obtain D-dimensional frame-level representations, which are then concatenated to form $\mathbf{H} \in \mathbb{R}^{L \times D}$. Next, the frame-level enrichment (FLE) sub-module enhances the frame representations globally by encoding the temporal context from different frames in the video and outputs *spatio-temporally enriched* frame-level representations $\mathbf{E} \in \mathbb{R}^{L \times D}$. These representations $\mathbf{E}$ are input to a temporal relationship modeling (TRM) module, which classifies the query video by matching its sub-sequences with support actions. Additionally, classifying intermediate representations $\mathbf{H}$ by introducing a query-class similarity classifier reinforces the learning of corresponding class-level information at different stages and aids in further improving the overall feature discriminability. Our framework is learned jointly using standard cross-entropy loss terms $\mathcal{L}_{TM}$ and $\mathcal{L}_{QC}$ on the class predictions from the TRM module and query-class similarity classifier, respectively. Next, we present our proposed spatio-temporal enrichment module.

3.2. Spatio-temporal Enrichment

The focus of our approach is the introduction of a spatio-temporal enrichment module that strives to enhance (i) local patch features spatially in an individual frame *and* (ii) global frame features temporally across frames in a video. The effective utilization of both spatial as well as temporal contextual information within a video enables improved class-specific feature discriminability before modeling the temporal relationships between query and support videos.

3.2.1 Enriching Local Patch Features

The patch features together in a frame encode its spatial information. Enhancing these features to encode the frame-level spatial context across all the patches in a frame is necessary to capture the appearance-based similarities as well as differences among the action classes. To this end, we introduce a patch-level enrichment (PLE) sub-module, which employs self-attention [27] to let the patch features attend to themselves by aggregating the congruent patch contexts. The PLE sub-module is illustrated in Fig. 3. Let $\mathbf{x}_i \in \mathbb{R}^{P^2 \times D}$ denote the latent features of P^2 patches in frame q_i ($i \in [1, L]$). Weights $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3 \in \mathbb{R}^{D \times D}$ project these latent features to obtain query-key-value triplets, given by

$$\mathbf{x}_i^q = \mathbf{x}_i \mathbf{W}_1, \quad \mathbf{x}_i^k = \mathbf{x}_i \mathbf{W}_2, \quad \mathbf{x}_i^v = \mathbf{x}_i \mathbf{W}_3. \tag{2}$$

While the *value* embedding persists the current status of a patch $p \in [1, P^2]$, the *query* and *key* vectors score the pairwise similarity between P^2 patches. These *value* embeddings are reweighted by corresponding normalized scores to obtain 'token-mixed' (attended) features $\boldsymbol{\alpha}_i$, given by

$$\boldsymbol{\alpha}_i = \eta \left(\frac{\mathbf{x}_i^q \mathbf{x}_i^{k\top}}{\sqrt{D}} \right) \mathbf{x}_i^v + \mathbf{x}_i, \tag{3}$$

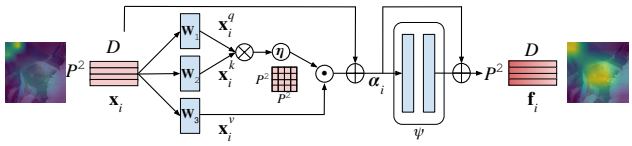


Figure 3. **Patch-level enrichment (PLE) sub-module.** Latent features $\mathbf{x}_i$ are projected by learnable weights $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{W}_3$ to form query-key-value triplets ($\mathbf{x}_i^q$, $\mathbf{x}_i^k$, $\mathbf{x}_i^v$). The value embeddings are re-weighted by the normalized pairwise scores between queries and keys, to obtain attended features $\boldsymbol{\alpha}_i$. A sub-network $\psi(\cdot)$ refines these $\boldsymbol{\alpha}_i$ to produce patch-level enriched features $\mathbf{f}_i$. Here, example attention maps before (on the left) and after (on the right) patch-level enrichment are shown. Best viewed zoomed in.

where η denotes softmax function. A sub-network $\psi(\cdot)$ then point-wise refines these attended features $\boldsymbol{\alpha}_i \in \mathbb{R}^{P^2 \times D}$ and outputs spatially enriched features $\mathbf{f}_i \in \mathbb{R}^{P^2 \times D}$, given by

$$\mathbf{f}_i = \psi(\boldsymbol{\alpha}_i) + \boldsymbol{\alpha}_i, \tag{4}$$

leading to an improved aggregation of the appearance-based action context across patches in a frame (see Fig. 5 row 3).

3.2.2 Enriching Global Frame Features

The local patch-level enrichment (PLE) described above aims to aggregate the spatial contexts locally within each frame of an action video. This enables focusing on relevant objects in a frame. However, it does not explicitly encode the temporal context and therefore struggles when encountered with object motion over time (see Fig. 5). Here, we proceed with the enrichment of temporal contexts globally across frames within a video by introducing a frame-level enrichment (FLE) sub-module comprising an MLP-mixer [25] layer. While self-attention is based on sample-dependent (input-specific) mixing guided by pairwise similarities between the tokens, the token-mixing in MLP-mixers assimilates the entire global receptive field through an input-independent and persistent relationship memory. Such a global assimilation of tokens enables the MLP-mixer to be better suited for enriching global frame representations. The FLE sub-module is shown in Fig. 4. For a frame q_i, let $\mathbf{h}_i \in \mathbb{R}^D$ denote the global representation obtained by spatially averaging the PLE output $\mathbf{f}_i \in \mathbb{R}^{P^2 \times D}$. The concatenated global representation $\mathbf{H} = [\mathbf{h}_1; \cdots; \mathbf{h}_L]^\top \in \mathbb{R}^{L \times D}$ for the entire video is then processed by the FLE sub-module. First, the frame tokens are mixed through a two-layer MLP $\mathbf{W}_t(\cdot)$ that is shared across channels (feature dimensions). This is followed by the token refinement of the intermediate features $\mathbf{H}_*$ by utilizing another two-layer MLP $\mathbf{W}_r(\cdot)$, which is shared across tokens. The two mixing operations in FLE are given by,

$$\mathbf{H}_* = \sigma(\mathbf{H}^\top \mathbf{W}_{t_1})\mathbf{W}_{t_2} + \mathbf{H}^\top, \tag{5}$$

$$\mathbf{E} = \sigma(\mathbf{H}_*^\top \mathbf{W}_{r_1})\mathbf{W}_{r_2} + \mathbf{H}_*^\top, \tag{6}$$

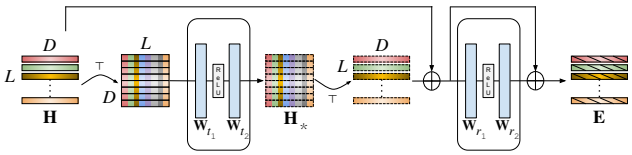


Figure 4. **Frame-level enrichment (FLE) sub-module.** The L frame tokens $\mathbf{h}_i$ of global representation $\mathbf{H}$ are first mixed through learnable weights $\mathbf{W}_{t_1}$ and $\mathbf{W}_{t_2}$ that are shared across the feature dimensions D, to obtain intermediate representations $\mathbf{H}_*$. This is then followed by individual token refinement using weights $\mathbf{W}_{r_1}$ and $\mathbf{W}_{r_2}$ that are shared across the L tokens, to obtain frame-level enriched features $\mathbf{E}$. Best viewed zoomed in.

where $\mathbf{E} \in \mathbb{R}^{L \times D}$ is the enriched feature, $\mathbf{W}_{t_1}, \mathbf{W}_{t_2} \in \mathbb{R}^{L \times L}$ and $\mathbf{W}_{r_1}, \mathbf{W}_{r_2} \in \mathbb{R}^{D \times D}$ are the learnable weights for token- and channel-mixing, respectively. Here, σ denotes the ReLU non-linearity. In particular, the token-mixing operation ensures that the frame representations interact together and imbibe the higher-order temporal relationships through the learnable weights $\mathbf{W}_{t_1}$ and $\mathbf{W}_{t_2}$. As a result, the FLE sub-module enhances the frame representations $\mathbf{h}_i$ temporally, with a global receptive field encompassing all the frames and produces temporally-enriched representations $\mathbf{e}_i$ for $i \in [1, L]$.

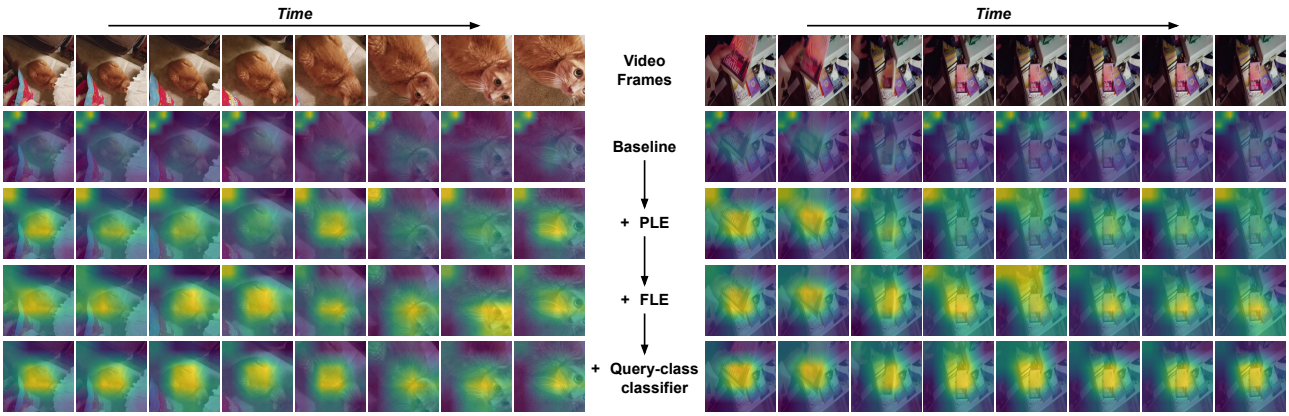
The enriched frame-level global representations $\mathbf{e}_i$ ($i \in [1, L]$) for the query and support videos are then input to the temporal relationship modeling (TRM) module, which models the temporal relationships between query and support actions. Within our framework, the TRM is a TRX (Eq. 1) built on a single cardinality $\Omega = \{2\}$, since our spatio-temporal enrichment module *learns* to model higher-order temporal representations without requiring multiple hand-crafted cardinality representations. Given the ground-truth labels $\mathbf{y} \in \mathbb{R}^C$, our framework is then learned end-to-end using the standard cross-entropy (CE) loss on the class probabilities $\hat{\mathbf{y}}_{TM} \in \mathbb{R}^C$ predicted by the TRM, given by

$$\mathcal{L}_{TM} = \mathbb{E}[\text{CE}(\hat{\mathbf{y}}_{TM}, \mathbf{y})]. \tag{7}$$

In summary, our spatio-temporal enrichment module leverages the advantages of local and global, sample-dependent and sample-agnostic enrichment mechanism to improve the aggregation of spatial as well as temporal contexts of actions. As a result, class-specific discriminative features are obtained along with the assimilation of higher-order temporal relationships in lower cardinality representations.

3.3. Query-class Similarity

As discussed above, the proposed framework comprising the feature extractor, spatio-temporal enrichment and temporal relationship modeling modules, is learned end-to-end with a CE loss on the output probabilities $\hat{\mathbf{y}}_{TM}$. However, learning to classify query video representations from intermediate layer outputs reinforces the model to look for class-



(*a*) Category: *Approaching Something with a Camera* (*b*) Category: *Dropping Something onto Something*

Figure 5. **The impact of progressively integrating our contributions** one at a time, from top (baseline) to bottom. The comparison is shown in terms of attention map visualizations measuring the activation magnitude of latent features for two examples from the SSv2 [11] test set. The baseline (second row) struggles to accurately capture the spatial as well as temporal contextual information. The integration of PLE sub-module (third row), which explicitly encodes spatial context, enables focus on relevant objects in a frame (second and third frame from the left in (a)). The integration of FLE sub-module (fourth row) further encodes the temporal context by consistently capturing the relevant object over time. For instance, while the context in fourth and sixth frame from the left in (a) are missed by PLE due to object motion, it is captured by the introduction of FLE. Lastly, integrating the query-class classifier further improves the attention on objects, leading to enhanced feature discriminability, *e.g.*, seventh and eighth frame from the left in (b) has improved attention on the object (book).

specific features at different stages in the pipeline. Consequently, such a multi-stage classification improves the feature discriminability, leading to better matching between query and support videos. To this end, we introduce a query-class similarity classifier on the patch-level enriched representations $\mathbf{h}_i$, $i \in [1, L]$. First, we obtain latent tuple representations $\mathbf{l}_t = [\mathbf{h}_{t_1}; \cdots; \mathbf{h}_{t_\omega}] \in \mathbb{R}^{\omega D}$ for tuples $t = (t_1, \cdots, t_\omega) \in \Pi_\omega$ in a video. They are then projected by $\mathbf{W}_{cls} \in \mathbb{R}^{\omega D \times D^{''}}$ to obtain $\mathbf{z}_t = \sigma(\mathbf{W}_{cls}^\top \mathbf{l}_t)$, where σ is the ReLU non-linearity. Then, for each $\mathbf{z}_t^Q$ in a query video Q, its highest similarity among all tuples in the K support videos for an action class c is computed. These scores for all the tuples in Q are aggregated to obtain the query-class similarity $M(Q, c)$ between the query and action c. With $\mathbf{z}_j^c$ representing a tuple $j \in [1, K \cdot |\Pi_\omega|]$ from the K support videos for an action c, the query-class similarity is given by

$$M(Q, c) = \sum_{\omega \in \Omega} \frac{1}{|\Pi_\omega|} \sum_{t \in \Pi_\omega} \max_j \phi(\mathbf{z}_t^Q, \mathbf{z}_j^c), \tag{8}$$

where $\phi(\cdot, \cdot)$ is a similarity function. Then, the C similarity scores are passed through *softmax* to obtain class probabilities $\hat{\mathbf{y}}_{QC} \in \mathbb{R}^C$ and trained with a CE loss given by

$$\mathcal{L}_{QC} = \mathbb{E}[\text{CE}(\hat{\mathbf{y}}_{QC}, \mathbf{y})]. \tag{9}$$

With λ as a hyper-weight, our STRM is trained using the joint formulation given by

$$\mathcal{L} = \mathcal{L}_{TM} + \lambda \mathcal{L}_{QC}. \tag{10}$$

Consequently, our proposed STRM, comprising a spatio-temporal enrichment module and an intermediate query-class similarity classifier, enhances feature discriminability (see Fig. 5) and leads to improved matching between queries and their support action classes.

4. Experiments

Datasets: Our approach is evaluated on four popular benchmarks: Something-Something V2 (SSv2) [11], Kinetics [4], HMDB51 [14] and UCF101 [23]. The SSv2 is crowd-sourced, challenging and has actions requiring temporal reasoning. For SSv2, we use the split with 64/12/24 action classes in training/validation/testing, given by [3]. A similar split with 64/12/24 action classes, as in [3, 34] is used for Kinetics. Furthermore, we evaluate on HMDB51 and UCF101 using the splits from [31]. The standard 5-way 5-shot evaluation is employed on all datasets and the average accuracy over 10,000 random test tasks is reported.

Implementation Details: As in [3, 19], a ResNet-50 [12], pretrained on ImageNet [6], is used as the feature extractor for $L = 8$ uniformly sampled frames of a video. With $D = 2{,}048$, an adaptive maxpooling reduces the spatial resolution to $P = 4$. All the learnable weights matrices in PLE and FLE are implemented as fully-connected (FC) layers. The sub-network $\psi(\cdot)$ in PLE is a 3-layer FC network with latent sizes set to 1,024. We set $D^{''} = 1{,}024$ for $\mathbf{W}_{cls}$. For the TRM, we employ $\Omega = \{2\}$ in Eq. 1 and set $D^{'} = 1{,}152$, as in [19]. The hyper-weight λ is set to 0.1. While 75,000 randomly sampled training episodes are used for SSv2 dataset with a learning rate of 10^{-3}, the smaller datasets are trained with a 10^{-4} learning rate. Our STRM framework is trained end-to-end using an SGD optimizer.

Table 1. **State-of-the-art comparison on four FS action recognition datasets**, in terms of classification accuracy. Our STRM outperforms existing FS action recognition methods on all four datasets. Importantly, for ResNet-50 backbone, STRM achieves an absolute gain of 3.5% over TRX [19] on the challenging SSv2 that comprises actions requiring temporal relationship reasoning.

Method	Backbone	Kinetics	SSv2	HMDB	UCF
CMN-J [34]	ResNet-50	78.9	-	-	-
TARN [2]	ResNet-50	78.5	-	-	-
ARN [31]	ResNet-50	82.4	-	60.6	83.1
OTAM [3]	ResNet-50	85.8	52.3	-	-
HF-AR [15]	ResNet-50	-	55.1	62.2	86.4
TRX [19]	ResNet-50	85.9	64.6	75.6	96.1
Ours:STRM	ResNet-50	**86.7**	**68.1**	**77.3**	**96.9**
TRX [19]	ViT	90.6	67.3	79.7	97.1
Ours:STRM	ViT	**91.2**	**70.2**	**81.3**	**98.1**

4.1. State-of-the-art Comparison

Tab. 1 shows the state-of-the-art comparison on four benchmarks for the standard 5-way 5-shot action recognition task. For fairness, only the approaches employing a 2D backbone for extracting per-frame features are compared in Tab. 1. On Kinetics, the recent works of OTAM [3] and TRX [19] achieve comparable classification accuracies of 85.8 and 85.9%. Our STRM performs favorably against existing methods by achieving an improved performance of 86.5%. On the more challenging SSv2 dataset comprising actions requiring temporal relational reasoning, OTAM and HF-AR [15] achieve 52.3% and 55.1%, while TRX obtains an accuracy of 64.6%, due to its temporal relationship modeling. Compared to the best existing approach of TRX, our STRM achieves a significant absolute gain of 3.5% on SSv2. Similarly, our STRM achieves improved performance on HMDB51 and UCF101, setting a new state-of-the-art on *all* four benchmarks. To further evaluate our contributions, we replace the ResNet-50 with ViT [8] as the backbone. Even with this stronger backbone, our STRM outperforms TRX on all datasets. In addition, our STRM achieves gains of 1.5% and 1.9% over TRX on SSv2, when employing 3D ResNet-50 and MViT [9] backbones. Note that the 3D ResNet-50 and MViT are pretrained on Kinetics400 [4] and are *not always* compatible with few-shot action datasets due to possible overlap of pretraining classes with novel classes. The consistent improvement of our STRM emphasizes the efficacy of enhancing spatio-temporal features, by integrating local (sample-dependent) patch-level and global (sample-agnostic) frame-level enrichment along with a query-class similarity classifier, for FS action recognition.

4.2. Ablation Study

Impact of the proposed contributions: Here, we systematically analyse the impact of our spatio-temporal enrichment module along with the query-class classifier. Note that our spatio-temporal enrichment module comprises PLE and FLE sub-modules. Fig. 6 (left) shows a performance comparison on SSv2, when integrating our two contributions (spatio-temporal enrichment module and the query-class classifier) in the baseline TRM. Note that the baseline TRM is a TRX [19] with cardinality $\Omega = \{2\}$. The baseline TRM achieves an FS action classification accuracy of 62.1% (red bar). Integrating our PLE in the baseline, for enriching the spatial context in the local patch-level features before temporal modeling, achieves an improved accuracy of 63.7% (orange bar). Similarly, enriching the temporal context alone in the global frame-level features through the integration of FLE (green bar) in TRM achieves a gain of 3.2%. Moreover, the joint integration of PLE and FLE (light blue bar) in the TRM further enhances the spatio-temporal contexts in the features, leading to an improved accuracy of 66.8%. Lastly, integrating the query-class classifier in our approach reinforces the learning of class-separable features at different stages and further enhances feature discriminability, thus, achieving a superior performance of 68.1%. The final STRM framework (blue bar) achieves an absolute gain of 6.0% over the baseline (red bar).

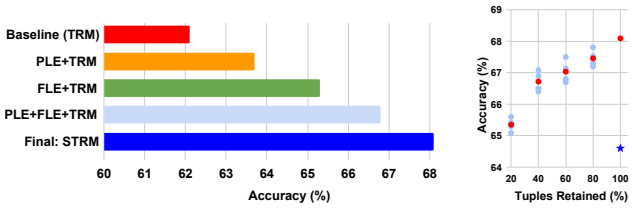


Figure 6. **(Left) Impact of integrating our contributions in the baseline** on SSv2. Individually integrating our PLE (orange bar) and FLE (green bar) into the baseline TRM results in improved performance. The joint integration (light blue bar) of PLE and FLE in the baseline enriches spatio-temporal features, leading to superior performance. Lastly, integrating our query-class classifier further enhances the feature discriminability. Our final STRM (blue bar) obtains an absolute gain of 6.0% over baseline. **(Right) Impact of varying #tuples in our STRM**. Multiple trials and mean performance of STRM are denoted by ● and ●, respectively. Since the feature discriminability is enhanced due to spatio-temporal enrichment, even with only 20% tuples retained, STRM ($\Omega = \{2\}$) performs favorably against TRX (denoted by ★) using $\Omega = \{2, 3\}$ and retaining all tuples. Best viewed zoomed in.

Impact of varying cardinalities: Tab. 2 shows the impact of varying the cardinalities considered for modeling temporal relationships in our STRM. The comparison is shown for Kinetics and SSv2. The number of tuples present in corresponding cardinality combinations is also shown. We observe that our STRM achieves optimal performance even at lower cardinalities. In particular, our STRM achieves the best performance on both datasets with $\Omega = \{2\}$. In contrast, TRX employing hand-crafted higher-order temporal representations requires $\Omega = \{2, 3\}$ to achieve its optimal performance of 64.6% on SSv2. Moreover, it is worth mentioning that our STRM is comparable to TRX in terms

Table 2. **Impact of varying the cardinalities for temporal relationships in our STRM** on Kinetics and SSv2. Here, we also show the number of tuples available in the corresponding cardinality combinations. Our STRM achieves best performance at a lower cardinality of $\Omega = \{2\}$, thereby mitigating the need of multiple TRM branches for different cardinalities.

Cardinalities (Ω)	$\{1\}$	$\{2\}$	$\{3\}$	$\{4\}$	$\{2,3\}$	$\{2,4\}$	$\{3,4\}$	$\{2,3,4\}$
#Tuples	-	28	56	70	84	98	126	154
Kinetics	86.2	**86.5**	86.0	85.3	85.9	86.1	85.7	86.1
SSv2	67.2	**68.1**	66.9	66.4	67.1	67.3	67.3	**68.1**

of compute, requiring only $\sim 4\%$ additional FLOPs. The superior performance of our approach over TRX at lower cardinality is due to the enhanced feature discriminability achieved through the spatio-temporal feature enrichment and the learning of higher-order temporal representations caused by token-mixing in our FLE sub-module.

Impact of varying tuples: Fig. 6 (right) shows the performance of our STRM approach on SSv2 when retaining different number of tuples for matching between query and support videos. We observe a marginal drop when the retained tuples are decreased. Moreover, even when retaining only 20% of the tuples at a lower cardinality ($\Omega = \{2\}$), our STRM achieves an accuracy of 65.4% and performs favorably against 64.6% of TRX, which relies on all the tuples from multiple cardinalities ($\Omega = \{2, 3\}$). This shows that our spatio-temporal enrichment module along with the query-class classifier enhances the feature discriminability while learning higher-order temporal representations in lower cardinalities itself. As a result, our STRM provides improved model flexibility, without requiring dedicated TRM branches for different cardinalities.

Comparison with different number of support samples: Fig. 7 compares STRM with the baseline and the TRX, when varying the number of support samples on SSv2. Here, we show K-shot ($K \leq 5$ and 10) classification. Our STRM achieves consistent improvement in performance, compared to both TRM and TRX on all K-shot settings. Specifically, our STRM excels in the extreme one-shot case as well as the 10-shot setting, where it effectively leverages larger support sets. Additional results are provided in the supplementary.

5. Relation to Prior Art

Several works have investigated the few-shot (FS) problem for image classification [1, 7, 10], object detection [13, 29], and segmentation [17]. While earlier approaches were either adaptation-based [18], generative [32], or metric-based [22, 28], recent works [7, 21] employ a combination of these. In the context of FS action recognition, [33, 34] employ memory networks for key-frame representations, whereas [2] aligns variable length query and support videos. Differently, [3] utilizes monotonic temporal ordering for enforcing temporal consistency between video pairs. The recent work of TRX [19] focuses on modeling the temporal

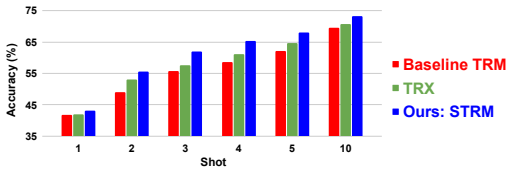


Figure 7. **Performance comparison when varying the number of support samples** in the SSv2 dataset. We show the comparison of our STRM with both TRM and TRX. STRM achieves superior performance compared to both TRM and TRX on all settings, including the challenging one-shot case. Furthermore, STRM effectively leverages larger support set in the 10-shot settings.

relationships by utilizing fixed higher-order temporal representations. Distinct from TRX, our STRM introduces a spatio-temporal enrichment module to produce spatio-temporally enriched features. The spatio-temporal enrichment module enriches features at local patch-level by employing a self-attention layer [20, 27, 30] as well as global frame-level by utilizing an MLP-mixer layer [24–26]. Our spatio-temporal enrichment also enables learning higher-order temporal representations at lower cardinalities. The proposed spatio-temporal enrichment module performs local patch-level enrichment using a self-attention layer as well as global frame-level enrichment by integrating a MLP-mixer, in a FS action recognition framework. Furthermore, we introduce a query-class classifier for learning to classify feature representations from intermediate layers.

6. Discussion

We proposed a FS action recognition framework, STRM, comprising spatio-temporal enrichment and temporal relationship modeling (TRM) modules along with a query-class similarity classifier. Our STRM leverages the advantages of combining local and global, sample-dependent and sample-agnostic enrichment mechanism for enhancing the spatio-temporal features, in addition to reinforcing class-separability of features at different stages. Consequently, this enhances the spatio-temporal feature discriminability and enables the learning of higher-order temporal relations even in lower cardinality representations. Our extensive ablations reveal the benefits of the proposed contributions, leading to state-of-the-art results on all benchmarks. A likely future direction, beyond the scope of current work, is to broaden the few-shot action recognition capability to generalize across varying domains.

Acknowledgements

This work was partially supported by VR starting grant (2016-05543), in addition to the compute support provided at the Swedish National Infrastructure for Computing (SNIC), partially funded by the Swedish Research Council through grant agreement 2018-05973.

References

[1] Peyman Bateni, Raghav Goyal, Vaden Masrani, Frank Wood, and Leonid Sigal. Improved few-shot visual classification. In *CVPR*, 2020. 8

[2] Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. *arXiv preprint arXiv:1907.09021*, 2019. 1, 7, 8

[3] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *CVPR*, 2020. 1, 6, 7, 8

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 1, 2, 6, 7

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *IJCV*, 2021. 1

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[7] Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: spatially-aware few-shot transfer. *arXiv preprint arXiv:2007.11498*, 2020. 1, 3, 8

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 7

[9] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021. 7

[10] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 8

[11] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 1, 2, 6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[13] Leonid Karlinsky, Joseph Shtok, Sivan Harary, Eli Schwartz, Amit Aides, Rogerio Feris, Raja Giryes, and Alex M Bronstein. Repmet: Representative-based metric learning for classification and few-shot object detection. In *CVPR*, 2019. 8

[14] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, 2011. 2, 6

[15] Neeraj Kumar and Siddhansh Narang. Few shot activity recognition using variational inference. *arXiv preprint arXiv:2108.08990*, 2021. 7

[16] Steinar Laenen and Luca Bertinetto. On episodes, prototypical networks, and few-shot learning. *arXiv preprint arXiv:2012.09831*, 2020. 3

[17] Weide Liu, Chi Zhang, Guosheng Lin, and Fayao Liu. Crnet: Cross-reference networks for few-shot segmentation. In *CVPR*, 2020. 8

[18] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 8

[19] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-relational crosstransformers for few-shot action recognition. In *CVPR*, 2021. 1, 2, 3, 6, 7, 8

[20] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. In *NeurIPS*, 2019. 8

[21] James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, and Richard E Turner. Fast and flexible multi-task classification using conditional neural adaptive processes. In *NeurIPS*, 2019. 8

[22] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017. 8

[23] 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. 2, 6

[24] Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. Synthesizer: Rethinking self-attention for transformer models. In *ICML*, 2021. 8

[25] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021. 5, 8

[26] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *arXiv preprint arXiv:2105.03404*, 2021. 8

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4, 8

[28] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NeurIPS*, 2016. 8

[29] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *arXiv preprint arXiv:2003.06957*, 2020. 8

[30] Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. Pay less attention with lightweight and dynamic convolutions. In *ICLR*, 2019. 8

[31] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip HS Torr, and Piotr Koniusz. Few-shot action recognition with permutation-invariant attention. In *ECCV*, 2020. 1, 6, 7

[32] Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. In *NeurIPS*, 2018. 8

[33] Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In *ECCV*, 2018. 8

[34] Linchao Zhu and Yi Yang. Label independent memory for semi-supervised few-shot video classification. *IEEE TPAMI*, 2020. 6, 7, 8