

Figure 1: Temporal relations alignment (prior works) vs. spatial relations alignment (our work). Compared to temporal inconsistency in (a), the spatial misalignment issue in (b) is more common.

ABSTRACT

Deep learning has achieved great success in video recognition, yet still struggles to recognize novel actions when faced with only a few examples. To tackle this challenge, few-shot action recognition methods have been proposed to transfer knowledge from a source dataset to a novel target dataset with only one or a few labeled videos. However, existing methods mainly focus on modeling the temporal relations between the query and support videos while ignoring the spatial relations. In this paper, we find that the spatial misalignment between objects also occurs in videos, notably more common than the temporal inconsistency. We are thus motivated to investigate the importance of spatial relations and propose a more accurate few-shot action recognition method that leverages both spatial and temporal information. Particularly, a novel Spatial Alignment Cross Transformer (SA-CT) which learns to re-adjust

* indicates equal contribution, # indicates corresponding authors.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 979-8-4007-0108-5/23/10...\$15.00 https://doi.org/10.1145/3581783.3612192 the spatial relations and incorporates the temporal information is contributed. Experiments reveal that, even without using any temporal information, the performance of SA-CT is comparable to temporal based methods on 3/4 benchmarks. To further incorporate the temporal information, we propose a simple yet effective Temporal Mixer module. The Temporal Mixer enhances the video representation and improves the performance of the full SA-CT model, achieving very competitive results. In this work, we also exploit large-scale pretrained models for few-shot action recognition, providing useful insights for this research direction.

CCS CONCEPTS

• Computing methodologies \rightarrow Activity recognition and understanding.

KEYWORDS

Action recognition; Few-shot learning; Meta learning; Cross attention; Spatial relation alignment

ACM Reference Format:

Yilun Zhang*, Yuqian Fu*, Xingjun Ma[#], Lizhe Qi, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang[#]. 2023. On the Importance of Spatial Relations for Few-shot Action Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3612192

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

1 INTRODUCTION

The massive amount and ever-growing video data demand automated video recognition techniques to analyze video content effectively for a multitude of multimedia applications. Powered by deep neural networks (DNNs), the performance of this task has been greatly improved in the past few years [5, 10, 23, 27, 40]. However, most of the mainstream video recognition methods rely heavily on the availability of massive labeled training data for improved performance. This is to assume that each target action has a large number of labeled examples. Unfortunately, such an assumption does not always hold in real-world applications. For example, it is unrealistic to collect videos for the actions of *drowning*, *failing*, etc. Recognizing actions in such a low-sample regime is thus of great importance—a challenging task also known as *few-shot action recognition* (FSAR).

Given a novel action with only a few labeled examples (also called support videos), the goal of FSAR is to classify the unseen query video according to the support videos. A body of work has been proposed to achieve accurate FASR from different perspectives [2, 3, 15, 26, 37, 44, 45, 52], amongst which exploiting the temporal relations between the query and the support videos is the latest and most effective approach. Representative methods include TAM [3], MTFAN [45], TRX [26], and STRM [37]. They all emphasize the importance of temporal relations and aim to match the support and query videos by aligning the temporal frames. However, in this paper, as depicted in Fig. 1, we reveal that the spatial misalignment of key objects also occurs in FSAR and is even more common than the temporal inconsistency, an important observation that has been neglected in previous works. This motivates us to re-examine the importance of spatial relations for few-shot action recognition, by answering the two questions: 1) Could spatial relations alone be sufficient for recognizing few-shot actions? and 2) Can temporal information be simply utilized as a boost to the final video representations?

To answer the first question, we propose a novel spatial crossattention (SCA) module to model the spatial relations between the query and support videos. Specifically, we first split the whole video frames into patches of equal size and then learn the similarities between the support and query patches via cross-attention. By matching those patches, objects of interest can be aligned even if they appear at different spatial locations. Through the aligned spatial relations, we turn the original support features into queryspecific support features, eliminating the negative effects caused by the spatial misalignment. Experimentally, we find that our SCA module alone without integrating any temporal information is already very competitive. On three out of four benchmarks, the SCA module achieves comparable or even better results than the current SOTAs. This indicates that the spatial relations play a central role in FSAR.

For the second question, we explore an alternative approach to exploit the temporal information, i.e., using it to boost the representations rather than temporally (and expensively) aligning the query and support videos as it did in prior works. Instead of designing sophisticated algorithms to match the temporal frames [2, 3], we explore a simple Temporal Mixer (TMixer) module for integrating the temporal information. We build the TMixer with several MLPs which is quite simple.

With the SCA module and the TMixer module, we are able to build a new state-of-the-art for FSAR.

Formally, a novel **S**patial Alignment **C**ross **T**ransformer **(SA-CT)** is proposed for few-shot action recognition. Our SA-CT is mainly composed of a feature extractor, a spatial cross-attention (SCA) module, and a Temporal Mixer (TMixer) module. Given a query video and a set of support videos, a feature extractor is first applied to extract the frame patch representations. The TMixer module is then employed to enrich the representations. The SCA module matches the patch representations between the query and support to construct query-specific support representations for classification. With SA-CT and its modules, we reveal that: 1) modeling the spatial relations alone can achieve comparable or even better results in most cases; 2) the temporal information can be effectively exploited as a temporal boost to the representations via a mixer module.

In addition, inspired by the recent success of large-scale pretrained models (LSPMs), we take a step further to explore several popular LSPMs including CLIP [28], DINO [4], and DeiT [39] for FSAR. Concretely, we leverage the backbones of these models as feature extractors to extract more enriched representations for videos. We empirically verify the benefit of using LSPMs for FSAR along with several useful insights. These explorations and analyses could help the community build more accurate FSAR models.

To summarize, we make the following contributions. **1**) We reveal the importance of spatial relations for FSAR, a long-overlooked aspect in the literature, and propose a novel spatial cross-attention (SCA) module to model the spatial relations for more accurate FSAR. **2**) We propose to use a Temporal Mixer (TMixer) module to integrate the temporal information into the representations as a complementary and boost to the spatial information. **3**) We combine the SCA and TMixer module into a unified Spatial Alignment Cross Transformer (SA-CT) architecture which achieves very competitive results for FSAR. **4**) We also provide the extensive exploration of LSPMs as more powerful feature extractors for FSAR.

2 RELATED WORK

Few-Shot Learning (FSL). Typical FSL methods can be roughly divided into three categories: model-based [25, 30], metric-based [20, 32, 34, 42] and optimization-based [11, 29]. Model-based methods aim to quickly update the parameters on a small number of samples through the design of the model structure, and directly establish the mapping function between the input instances and the predictions. Metric-based methods measure the distances (e.g., cosine similarity) between the samples in the support set and query set. More recent FSL methods [8, 12, 14, 35, 36, 47-49, 51, 53] include HyperTransformer [51], Meta-FDMixup [12], STraTA [43], CrossTransformers [8], MetaQDA [49], PMF [19], and StyleAdv [14]. Among them, CrossTransformers [8] is the most related work to ours, which uses the attention mechanism to find spatial correspondence between the query and the labeled images. However, CrossTransformers was proposed for image few-shot learning, while we tackles the more challenging problem of video few-shot learning.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

Few-Shot Action Recognition (FSAR). FSAR aims to recognize unseen videos with only a few labeled samples. Prior works have made certain progresses via including compound memory network [52] for optimal video representations, synthesizing additional examples for novel categories [22], leveraging synthetic videos as data augmentation [13], and introducing extra multimodal information [15]. More recent works focus on utilizing temporal information. Inspired by the text sequence matching task, TARN [2] regards videos as segment-level sequence data and matches the query with the support videos. OTAM [3] aligns query and support videos temporally by calculating frame similarities. In order to align subsequences of actions at different speeds, TRX [26] constructs video representations from ordered tuples of varying numbers of frames. MTFAN [45] explored the task-specific motion modulation and the multi-level temporal fragment alignment. HyRSM [44] brought up hybrid relation module and set matching metric. By adding spatial and temporal enrichment module to TRX [26], STRM [37] achieves impressive performance in FSAR.

Cross-Attention for Few-Shot Learning. One core challenge of few-shot learning lies in matching the support and query instances. Following this, several methods [8, 18, 26, 37, 50] have been proposed to explore cross-attention for improved alignment of the image/video instances. Among them, TRX [26] and STRM [37] are the two most related works to us. Specifically, TRX shows its potential in matching actions at different speeds. Based on TRX, STRM reaches very competitive performance by further adding a self-attention module on spatial patchs. Different from TRX and STRM which both tackle FSAR from a temporal perspective, our work investigates the importance of spatial relations to FSAR and proposes to align the spatial objects between videos to achieve more accurate FSAR. Note that STRM also handles the spatial relation, but via self-attention applied on patches within a single video. By contrast, we apply cross-attention to interact patches between the support and query videos.

Large-Scale Pretrained Models (LSPMs). Previous works have shown the significant impacts of LSPMs to various downstream tasks [4, 9, 24, 28, 39]. Those LSPMs exceed the CNN network in a number of vision tasks. A recent work [19] shows that a simple transformer-based pipeline can boost the performance of FSL. However, applying LSPMs for FSAR is still underexplored. Existing FSAR methods [3, 15, 26, 37] mainly focus on a few standard CNN architectures (e.g., ResNet-50 pretrained on ImageNet). In this paper, we explore whether FSAR can also benefit from LSPMs and to what extent LSPMs can boost the performance of our model.

3 PROPOSED METHOD

3.1 **Problem Formulation**

Given *C* action classes each containing only *K* (a small number like 5) labelled instances as the 'support set', the task of FSAR is to classify an unlabelled query video into one of the classes in the 'support set'. Following the episodic training paradigm in prior works [3, 11, 42, 46], we use episodic training in which few-shot tasks are sampled randomly from the training set. In each episode, we learn a *C*-way *K*-shot classification task. We denote a query video of *L* frames as $Q = \{q_1, \dots, q_L\}$. For each class $c \in \{1, \dots, C\}$,

we denote the support videos of this class as S^c . Specifically, S^c contains *K* videos, for the k^{th} video, we have $S_k^c = \{s_{k1}^c, \dots, s_{kL}^c\}$.

3.2 Method Overview

Motivated by the observation that the spatial misalignment of key objects is oftentimes more severe than the temporal inconsistency, we introduce a novel Spatial Alignment Cross Transformer (SA-CT) to better handle the spatial misalignment by readjusting the spatial relations between two videos. The overall architecture is shown in Fig. 2. First, a feature extractor is leveraged to encode the video frames in the support and query sets. The frame features are then passed through the TMixer module to fuse higher-order temporal information as well as to reduce the number of video frames. Next, the SCA module is employed to match the spatial patches between the query and support set, and construct the query-specific prototypes for classification. Finally, distances are passed as logits for training and inference steps.

3.3 Spatial Cross-Attention Module (SCA)

Cross-attention mechanism, initially introduced in [41], has shown its ability in aligning images and videos [3, 15, 26, 37]. We employ this mechanism in our SA-CT to align the spatial objects in the query and support set. The procedure is illustrated in Fig. 3. Specifically, we define the query feature of the i^{th} ($i \in [1, L]$) frame at the spatial position p as:

$$qf_{ip} = \left[\Phi\left(q_i\right)_p + \text{CPE}\left(\Phi\left(q_i\right)_p\right)\right],\tag{1}$$

where $\Phi : \mathbb{R}^{H \times W \times 3} \mapsto \mathbb{R}^{P^2 \times D}$ represents a feature extractor that extracts the latent features of P^2 patches (i.e., spatial position $p \in [0, P^2]$), and $\text{CPE}(\cdot)$ is a conditional positional encoding of a frame feature [6]. Similarly, the feature at spatial position *m* of the *i*th $(i \in [1, L])$ frame of video *k* in the support set of class *c* is:

$$sf_{ikm}^{c} = \left[\Phi\left(s_{ik}^{c}\right)_{m} + \text{CPE}\left(\Phi\left(s_{ik}^{c}\right)_{m}\right)\right].$$
 (2)

After generating the features of the query and support videos, the similarity between the query and support at the i^{th} ($i \in [1, L]$) frame can be calculated as:

$$a_{ikmp}^{c} = LN\left(W_{q} \cdot sf_{ikm}^{c}\right) \cdot LN\left(W_{k} \cdot qf_{ip}\right),\tag{3}$$

where $LN(\cdot)$ is the standard layer normalization [1], and W_q , W_k are learnable projections of the features into the query and key embeddings used in the attention mechanism [41].

For spatial alignment, each of the support videos is fully utilized, i.e., patches in the query video are matched with those at different locations of all the videos in support class *c*. The attention map can thus be derived by applying the Softmax operation along patches of videos in a support class:

$$\tilde{a}_{ikmp}^{c} = \frac{\exp\left(a_{ikmp}^{c}/\tau\right)}{\sum_{l,n} \exp\left(a_{ilnp}^{c}/\tau\right)}, \tau = \sqrt{d_k}.$$
(4)

The above attention map represents the correspondences between different spatial locations in the query and support videos. This allows the module to readjust the support features into a MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

Zhang et al.



Figure 2: The architecture of our proposed Spatial Alignment Cross Transformer (SA-CT), illustrated in a 1-way 2-shot case. It consists of a 1) feature extractor that extracts the features for both the support set and the query; 2) a TMixer module that integrates the higher-order temporal information; and 3) a SCA module that aligns the spatial patches between the query and support set, and construct the query-specific prototypes for classification.



Figure 3: The architecture of the SCA module. Each corresponding frame of patches in the query and support videos are first passed through the Linear weights to compute their cross attention. A query-specific prototype is then constructed based on the attention map. Finally, a mean distance value is computed by averaging over the distances of all video frames.

query-specific prototype for better spatial alignment as follows:

$$\mathbf{t}_{ip}^{c} = \sum_{km} \tilde{a}_{ikmp}^{c} \cdot \left(\mathbf{W}_{\mathbf{v}} \cdot \Phi \left(s_{ik}^{c} \right)_{m} \right), \tag{5}$$

where, W_v represents the attention value weights.

Finally, the module calculates the distances between the queryspecific prototypes and the query using squared Euclidean distance, and parses the distances as logits to represent the distribution over the classes:

$$d\left(Q,S^{c}\right) = \frac{1}{P^{2}} \sum_{p} \left\| \frac{1}{L^{2}} \sum_{i} \left(t_{ip}^{c} - \left(\mathbf{W}_{\mathbf{v}} \cdot \Phi\left(q_{i}\right)_{p} \right) \right) \right\|_{2}^{2}.$$
 (6)

The same value weight W_v is applied to both the query and support features to ensure that the distances calculated above could measure the similarities between the query and support videos. We also set the query weight W_q and key weight w_k to be the same, which could maximize the attention value for the corresponding spatial locations.

3.4 Temporal Mixer Module (TMixer)

The SCA module enables our SA-CT to align objects in the query and support videos. In addition to spatial relations, we explore a TMixer module that modifies two MLP-mixer [38] layers to integrate the temporal information in a more efficient manner. A standard MLP-Mixer consists of two types of MLP layers: channelmixing MLPs and token-mixing MLPs. The main purpose of channelmixing MLPs is to facilitate communication between different channels, which inspires us to integrate the inter-frame relations with a global reception field. In our TMixer module, we regard frames of a video as different channels. Frame features are thus able to interact with each other and be enriched globally with high-order temporal information. An illustration of this module is depicted in Fig. 4. Concretely, consider the feature f_i obtained by a backbone at the i_{th} ($i \in [1, L]$) frame, then the concatenated feature representations



Figure 4: The architecture of the TMixer module. The frame representations are passed through four MLPs, each of which consists two layers. The first two MLPs integrate the temporal information by allowing the frames to interact with each other. The last two MLPs are employed to reduce the number of frames, thus decreasing the computational cost and accelerating the model.

of an entire video are denoted as $\mathbf{F} = [\mathbf{f}_1; \cdots; \mathbf{f}_L] \in \mathbb{R}^{L \times P^2 \times D}$. We employ two MLPs, to interact frames of a single video:

$$\mathbf{U}_{i,*,*} = \mathbf{F}_{i,*,*} + \mathbf{W}_2 \cdot \sigma \left(\mathbf{W}_1 \cdot \mathbf{F}_{i,*,*} \right), i = 1 \dots L,$$
(7)

$$\mathbf{V}_{*,*,j} = \mathbf{U}_{*,*,j} + \mathbf{W}_4 \cdot \sigma \left(\mathbf{W}_3 \cdot \mathbf{U}_{*,*,j} \right), j = 1 \dots D,$$
(8)

where, σ denotes the ReLU non-linearity, and $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{L \times L}, \mathbf{W}_3, \mathbf{W}_4 \in \mathbb{R}^{D \times D}$ are all learnable weights.

We have enriched the video features globally following the idea of frame mixing, and now we further extend this idea to reducing the frames. Concretely, the following two MLPs are used to reduce the frames and accelerate the aforementioned SCA modules:

$$\mathbf{Y}_{*,*,*} = \mathbf{W}_6 \cdot \sigma \left(\mathbf{W}_5 \cdot \mathbf{V}_{i,*,*} \right), i = 1 \dots L,$$
(9)

$$\mathbf{Z}_{*,*,j} = \mathbf{Y}_{*,*,j} + \mathbf{W}_8 \cdot \sigma \left(\mathbf{W}_7 \cdot \mathbf{Y}_{*,*,j} \right), j = 1 \dots D, \tag{10}$$

where, σ denotes the ReLU non-linearity, and $\mathbf{W}_5 \in \mathbb{R}^{L \times L/2} \mathbf{W}_6 \in \mathbb{R}^{L/2 \times L/2}$, \mathbf{W}_7 , $\mathbf{W}_8 \in \mathbb{R}^{D \times D}$ are all learnable weights. After passing through these two MLPs, *L* frames in each video are reduced to *L*/2 frames, we will empirically verify the effectiveness of this operation in Sec. 4.4.

Note that this module is applied on both the query and the support features before passing through the SCA module. As such, features of one single frame are enabled to incorporate high-order temporal information of the whole video.

3.5 Relation to Prior Works

The recent works TRX [26] and STRM [37] both explore using crossattention to address FSAR tasks. TRX [26] focuses on aligning subsequences of videos. On the other hand, STRM [37] incorporates spatial information by employing self-attention on individual video frames, enriching them locally, and then passing these enriched frames into a modified TRX module (named TRM) for recognition. Despite this, the core concept of STRM remains similar to TRX, as it still involves video sub-sequences matching. In contrast, our approach differs from TRX and STRM. We utilize cross-attention to match patches in both support and query video frames, enabling an interaction between all the support and query patches. This enables our proposed SA-CT model to spatially align videos effectively.

4 EXPERIMENTS

4.1 Experimental Setup

Datasets. We evaluate our method on four action recognition datasets, including HMDB51 [21], UCF101 [33], Kinetics [5], and Something-Something V2 (SSv2) [16]. UCF101 dataset contains 101 action categories of over 13,320 short trimmed videos; HMDB51 dataset has 6,849 videos from 51 action categories; Kinetics contains 400 types of human actions and each video lasts around 10 seconds; SSv2 is the largest dataset containing over 220k videos. One common characteristic of the UCF101, HMDB51, and Kinetics datasets is that the semantic concepts (e.g., objects and backgrounds) are more related to the action categories. However, for SSv2, a large proportion of the categories are more related to the temporal information [31]. As for the specific splits of training, validation, and testing sets, for UCF101 and HMDB51, we use the splits as in ARN [46]. For Kinetics, we follow the splits proposed in CMN [52], in which 100 classes are selected to split into 64/12/24 classes for train/val/test, respectively. For SSv2, we follow the same splits as in OTAM [3], which also define 64/12/24 action classes for train/val/test sets.

Implementation Details. For fair comparison, following previous works [3, 15, 26, 52], we adopt the ResNet-50 [17] pretrained on ImageNet [7] as our backbone. We remove the last two layers of ResNet-50 to extract feature maps of size $7 \times 7 \times 2048$ for video frames. The extracted feature maps can support our cross-attention operation on patches. For each video (in both the query and support set), we uniformly sample 8 frames, i.e., L = 8. We first re-scale the height of the frames to 256 and then crop the frames to 224×224 . Common data augmentations like random cropping and horizontal flipping are also used during model training. The SGD is used as the optimizer for the meta-train stage. For UCF101, HMDB51, and Kinetics, the learning rate is set to 0.0005; while for SSv2, we use a learning rate of 0.005 due to its larger number of videos. During training, the validate set is used to determine the hyper-parameters with the best are then adopted for testing. For testing, we randomly select 10,000 meta-tasks from the test set and report the average accuracy as the final performance metric.

4.2 Comparison with SOTAs

To show the effectiveness of our SA-CT, we first compare our method with the most representative and SOTA methods, including the CMN [52], TRAN [2], Embodied [13], ARN [46], OTAM [3], AmeFu-Net [15], TRX [26], and STRM [37]. Our main results for 5-way 1-shot and 5-way 5-shot FSAR tasks are reported in Tab. 1.

Results without the TMixer. We first compare our SA-CT when the TMixer module is removed so as to verifies the importance of

Method	Backbone	UCF101 HMDB51		SSv2		Kine	tics		
		1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot
CMN [52]	ResNet-50	-	-	-	-	-	-	60.5	78.9
TARN [2]	ResNet-50	-	-	-	-	-	-	64.8	78.5
Embodied [13]	ResNet-50	-	-	-	-	-	-	67.8	85.0
ARN [46]	ResNet-50	66.3	83.1	45.5	60.6	-	-	63.7	82.4
OTAM [3]	ResNet-50	-	-	-	-	42.8	52.3	73.0	85.8
AmeFu-Net [15]	ResNet-50	85.1	95.5	60.2	75.5	-	-	74.1	86.8
TRX [26]	ResNet-50	-	96.1	-	75.6	-	64.6	-	85.9
STRM [37]	ResNet-50	-	96.9	-	77.3	-	68.1	-	86.7
SA-CT w/o TMixer (ours)	ResNet-50	-	96.4	-	77.8	-	61.4	-	86.4
SA-CT (ours)	ResNet-50	85.4	96.4	60.4	78.3	48.9	69.1	71.9	87.1
SA-CT (ours)	ViT-base	-	98.0	-	81.6	-	66.3	-	91.2

Table 1: Comparisons with SOTAs. We hightlight that: 1) our "SA-CT w/o TMixer" that without any temporal information used outperforms the TRX and is comparable to the STRM on UCF101, HMDB51, and Kinetics. This reveals the importance of the spatial relation; 2) Our full SA-CT model improves the "SA-CT w/o TMixer" and achieves very competitive results. This indicates that the temporal information could be well utilized via a simple temporal module.

spatial relation (captured by the SCA module). As shown in "SA-CT w/o TMixer" in Tab. 1, surprisingly, our SA-CT without utilizing any temporal information (i.e., "SA-CT w/o TMixer (ours)") without utilizing any temporal information achieves very competitive results to the current SOTAs on three out of four datasets. Particularly, compared with TRX that only tackles the temporal misalignment, our "SA-CT w/o TMixer" performs better on UCF101, HMDB51, and Kinetics datasets. This indicates that, on relatively common video datasets, aligning the spatial semantic concepts is more important than fixing the temporal inconsistencies. Even when compared with the STRM method which upgrades TRX by using the spatial information as a supplement, our results are still comparable on this three datasets. On HMDB51, we can even surpass STRM by 0.5%. Note that AmeFu-Net also reaches very competitive results, however, AmeFu-Net additionally introduces the depth modality while we only require the RGB frames. These results confirm the importance of modeling spatial relations for FSAR. Our "SA-CT w/o TMixer" does not work well on SSv2, which we conjecture is caused by the complicated temporal information contained in SSv2 videos. Arguably, such cases are relatively less common in real-world applications. The limitation on SSv2 in turn motivates us to incorporate the temporal information into our SA-CT by the TMixer to in turn boost the spatial relation.

Results of the Full SA-CT. Equipped with the TMixer, our full SA-CT model outperforms the baselines by a considerable margin consistently across different datasets and settings. Specifically, **1**) under the 5-way 1-shot setting, SA-CT achieves an accuracy of 85.4%, 60.4%, 48.9%, and 71.9% on the UCF101, HMDB51, SSv2, and Kinetics, respectively. We improve the ARN model by up to 19.1%, 14.9%, and 8.2% respectively on UCF101, HMDB51, and Kinetics datasets. The superiority in such a low data regime indicates the efficiency of our SA-CT approach in capturing and aligning the spatial relations between the query and support videos. **2**) Under the 5-way 5-shot setting, our SA-CT also achieves good results. On HMDB51, SSv2, and Kinetics, we outperform the best baseline STRM by 1.0%, 1.0%, and 0.4%, respectively. Though the performance improvement

may *numerically* seems minor when compared with the improvement achieved under the 1-shot setting, we highlight that both TRX and STRM are well-developed and sophisticated FSAR algorithms that work really well under the 5-shot setting, and it is notably hard to exceed the two methods under this relatively data sufficient regime. **3)** By employing the ViT-base architecture, we are able to achieve new state-of-the-art (SOTA) results on three benchmark datasets: UCF101, HMDB51, and Kinetics. The obtained accuracies for these datasets are 98.0%, 81.6%, and 91.2%, respectively. For a comprehensive analysis of various backbones and their impact on performance, please refer to Sec. 4.3.

4.3 Learning From LSPMs.

In this section, we investigate the benefit of using LSPMs for FSAR. Besides the ResNet-50 that has been widely used for FSAR [3, 15, 26, 37], here we first explore different pretrained vision transformers as the feature extractors. We build our SA-CT method upon those extractors and use different LSPMs, including DINO [4], CLIP [28], DeiT [39], and supervised training method (abbreviated as "SL") to initialize the backbones. We conduct experiments on both fixing and freeing the extractor parameters. In addition, to evaluate the representations extracted by different extractors, we further introduce the ProtoNet (PN) [32] as a FSL classifier to build simple baselines. Concretely, PN is a non-prametric FSL method which classifies the query actions by ranking the feature similarities between instances. Thus, we form the "PN-FSAR" by simply averaging the frames features and then feeding the averaged features into PN. The 5-way 5-shot results are reported in Tab. 2. From the results, we summarize the following observations:

• A new SOTA (highlighted in green) can be achieved by building our SA-CT upon the ViT-base (SL/IN21K). On UCF, HMDB, and Kinetics, the new SOTA accuracy reaches 98.0%, 81.6%, and 91.2%, respectively. Compared to our prior SA-CT (ResNet-50)(SL/IN1K) which is highlighted in orange, the performances on UCF, HMDB, and Kinetics are improved by up to 1.6%, 3.3% and 4.1%, respectively. These results confirm that FSAR task can indeed be improved by using LSPMs.

Extractor	Pretrain	Fix	Method	UCF	HMDB	SSv2	Kinetics
ResNet-50	SL/IN1K	\checkmark	PN-FSAR	88.8	61.1	39.9	76.7
ResNet-50	SL/IN1K	\checkmark	SA-CT	94.2	73.7	59.5	85.3
ResNet-50	SL/IN1K	-	SA-CT	96.4	78.3	69.1	87.1
ResNet-50	DINO/IN1K	\checkmark	PN-FSAR	91.3	66.9	40.3	78.2
ResNet-50	DINO/IN1K	\checkmark	SA-CT	93.4	75.2	57.6	84.3
ResNet-50	DINO/IN1K	-	SA-CT	95.0	75.4	65.2	85.8
ResNet-50	CLIP/YFCC	\checkmark	PN-FSAR	91.3	70.6	36.5	82.2
ResNet-50	CLIP/YFCC	\checkmark	SA-CT	96.0	79.0	61.2	87.8
ViT-small	DINO/IN1K	\checkmark	PN-FSAR	93.2	69.0	42.9	82.8
ViT-small	DINO/IN1K	\checkmark	SA-CT	93.3	70.6	63.2	78.5
ViT-small	DINO/IN1K	-	SA-CT	95.4	72.2	61.0	82.8
ViT-small	DeiT/IN1K	\checkmark	PN-FSAR	91.8	64.2	36.8	82.7
ViT-small	DeiT/IN1K	\checkmark	SA-CT	93.0	72.1	58.1	83.7
ViT-small	DeiT/IN1K	-	SA-CT	95.7	77.8	66.1	86.1
ViT-base	SL/IN21K	\checkmark	PN-FSAR	96.7	76.3	41.4	90.3
ViT-base	SL/IN21K	\checkmark	SA-CT	96.3	77.6	51.8	88.6
ViT-base	SL/IN21K	-	SA-CT	98.0	81.6	66.3	91.2
ViT-base	DINO/IN1K	\checkmark	PN-FSAR	94.4	70.5	43.2	84.7
ViT-base	DINO/IN1K	\checkmark	SA-CT	94.6	72.6	63.3	81.2
ViT-base	DeiT/IN1K	\checkmark	PN-FSAR	92.8	67.0	35.2	82.0
ViT-base	DeiT/IN1K	\checkmark	SA-CT	95.1	75.1	55.5	85.2

Table 2: Learning from LSPMs. The 5-way 5-shot results are reported here. The "PN-FSAR" denotes the simple baseline that takes the non-parametric protonet as the FSL classifier. It basically evaluates the ability of LSPMs. Both the ResNet-50 and ViT are investigated with different pre-training methods. Compared to our method (based on typical ResNet-50 (IN1K), highlighted in orange), a new SOTA (highlighted in green) is achieved by employing the ViT-base (ImageNet21k).

- ViT has advantages over the ResNet-50 on extracting richer representations for video frames. Generally, with the same "DINO/IN1K" as pretraining, comparing the results of PN-FSAR, the ViT-base performs the best, followed by the ViTsmall, while ResNet-50 performs the worst.
- Supervised training vs. self-supervised training. Overall, though the new SOTA is achieved by the ViT-base pretrained under supervised learning (SL), we notice that the self-supervised learning (SSL) pretraining methods (e.g. DINO) also achieves good results. Specifically, taking ResNet-50 as the extractor, the SSL based "DINO/IN1K" is competitive to "SL/IN1K". For the ResNet-50 with PN-FSAR, the "DINO/IN1K" even outperforms the "SL/IN1K".
- Finetuning the extractor improves performance. Comparing the results of fixing the extractor or not, we find that finetuning plays an important role in improving FSAR. Considering the potential data shifts between the pretraining and testing datasets, such an improvement is somewhat expected. That is, finetuning the base model on the specific testing data can help alleviate distributional shifts and improve performance.
- *SSv2 is an exception.* Interestingly, among all the testing datasets, the UCF, HMDB, and Kinetics all benefit from LSPMs more or less, while SSv2 is the only exception. This reflects different properties of the testing datasets, i.e., compared to other datasets, SSv2 has a higher demand for temporal information [31]. Unfortunately, the LSPMs were all pretrained on images thus become less effective on datasets like SSv2. This calls for dedicated LSPMs for video tasks.

4.4 Ablation Study

The impact of TMixer. In Fig. 5, we demonstrate that the TMixer module significantly improves SA-CT performance. On



Figure 5: The impact of TMixer. The simple TMixer module improves the SA-CT steadily.

HMDB51 and Kinetics, accuracy increases from 77.8% and 78.3% to 87.1% and 87.1%, respectively. On SSv2, TMixer improves the performance from 61.4% to 69.1%, with a 7.7% gain. These results confirm the effective utilization of temporal information using a simple MLP-based module. TMixer also adds MLP3 and MLP4, reducing video frames from 8 to 4. This decreases the total Multi-Adds of SAC by up to 38.5% (from 5.48G to 3.37G), as shown in Tab. 3. The additional two MLPs contribute only 419.64M Multi-Adds. This demonstrates that TMixer is a simple and efficient module for utilizing temporal information and reducing computational costs.

Frames	Total Multi-Adds of SAC	Total Multi-Adds of TMixer
8	5.48G	419.64M
4	3.37G	839.28M

Table 3: TMixer accelerates the model. The Multi-Adds of SAC is decreased by 38.5% (from 5.48G to 3.37G).

Varying the number of patches. Here we study the impact of varying number of patches on HMDB51 dataset. In Fig. 7, the horizontal axis represents the number of patches in each row (e.g. 7 means each row has 7 patches, and one video frame has $7 \times 7 = 49$



Figure 6: Visualizations of our spatial alignment. Four examples are given. The horizontal and vertical axes represent the support and query patches, respectively. The brighter the color, the higher the similarity. We downsample the total patches from $7 \times 7 = 49$ patches to $3 \times 3 = 9$ patches for better visualization. Results show that our SA-CT successfully aligns the related regions of the support and query videos.



Figure 7: Ablation study on the number of patches. We conduct experiments on HMDB51 with the number of patches in a row set varying from 1 to 7. SA-CT achieves the best result when the frames are split into $7 \times 7 = 49$ patches.

patches), while the vertical axis represents the accuracy of the model. Generally, we observe that the accuracy of the performance improves as the number of patches increases with exceptions on 3 patches and 6 patches. Critically, our SA-CT achieves the best performance when the number of patches is set as 7; and the performances are relatively poor when the number of patches is smaller than 4. This indicates that with more fine-grained patches, the benefits of our SCA module are maximized.

4.5 Visualization

In this section, we visualize four examples of correcting the spatial misalignment in Fig. 6 to help understand the working mechanism of our SA-CT. Each example consists of a query frame image and a few support frame images, and the corresponding attention maps.

We downsample the total patches from $7 \times 7 = 49$ patches to $3 \times 3 = 9$ patches for better visualization. The vertical axis represent the query patches (3×3 grid, flatten to 9 patches), and horizontal axis represent support patches (3×3 grid, flatten to 9 patches). The brighter the color, the higher the similarity. The visualized examples show clearly the advantage of our SA-CT in aligning the objects of interest from different locations. Taking "playing ukulele" as an example, our model successfully matches the "ukulele" patch even when it appears at different spatial locations of the two videos. These results well indicate the effectiveness of our SCA module.

5 CONCLUSION

In this paper, we re-examined the role of spatial relations for Few-Shot Action Recognition (FSAR), and experimentally revealed its importance for accurate FSAR via a proposed Spatial Cross-Attention (SCA) module. With SCA, we introduced a novel **S**patial **A**lignment Cross Transformer (**SA-CT**) which better handles the spatial misalignment by re-adjusting the spatial relations between two videos. Further equipped with a simple but effective Temporal Mixer module, our SA-CT achieves state-of-the-art results across different benchmark datasets and settings. We also conducted extensive experiments to explore the potential of LSPMs for FSAR. Our work contributes to the field with a new SOTA method and useful understanding for future research.

ACKNOWLEDGMENTS

This project is in part supported by NSFC under Grant No. 62032006 and No. 62072116.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

REFERENCES

- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. arXiv preprint arXiv:1607.06450 (2016).
- [2] Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. 2019. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. arXiv preprint arXiv:1907.09021 (2019).
- [3] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. 2020. Few-shot video classification via temporal alignment. In CVPR.
- [4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In ICCV.
- [5] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In CVPR.
- [6] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. 2021. Conditional positional encodings for vision transformers. arXiv preprint arXiv:2102.10882 (2021).
- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In CVPR.
- [8] Carl Doersch, Ankush Gupta, and Andrew Zisserman. 2020. Crosstransformers: spatially-aware few-shot transfer. *NeurIPS* (2020).
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.
- [10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In ICCV.
- [11] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic metalearning for fast adaptation of deep networks. In *ICML*.
- [12] Yuqian Fu, Yanwei Fu, and Yu-Gang Jiang. 2021. Meta-fdmixup: Cross-domain few-shot learning guided by labeled target data. In ACM Multimedia.
- [13] Yuqian Fu, Chengrong Wang, Yanwei Fu, Yu-Xiong Wang, Cong Bai, Xiangyang Xue, and Yu-Gang Jiang. 2019. Embodied one-shot video recognition: Learning from actions of a virtual embodied agent. In ACM Multimedia.
- [14] Yuqian Fu, Yu Xie, Yanwei Fu, and Yu-Gang Jiang. [n. d.]. StyleAdv: Meta Style Adversarial Training for Cross-Domain Few-Shot Learning. In CVPR.
- [15] Yuqian Fu, Li Zhang, Junke Wang, Yanwei Fu, and Yu-Gang Jiang. 2020. Depth guided adaptive meta-fusion network for few-shot video recognition. In ACM Multimedia.
- [16] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. 2017. The" something something" video database for learning and evaluating visual common sense. In *ICCV*.
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In CVPR.
- [18] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. 2019. Cross attention network for few-shot classification. *NeurIPS* (2019).
- [19] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M Hospedales. 2022. Pushing the Limits of Simple Pipelines for Few-Shot Learning: External Data and Fine-Tuning Make a Difference. In CVPR.
- [20] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*.
- [21] Hildegard Kuehne, Hueihan Jhuang, Estibaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: a large video database for human motion recognition. In *ICCV*.
- [22] Sai Kumar Dwivedi, Vikram Gupta, Rahul Mitra, Shuaib Ahmed, and Arjun Jain. 2019. Protogan: Towards few shot learning for action recognition. In *ICCV Workshops*.
- [23] Ji Lin, Chuang Gan, and Song Han. 2019. Tsm: Temporal shift module for efficient video understanding. In *ICCV*.
- [24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*.
- [25] Tsendsuren Munkhdalai and Hong Yu. 2017. Meta networks. In ICML.
- [26] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. 2021. Temporal-relational crosstransformers for few-shot action recognition. In CVPR.
- [27] Zhaofan Qiu, Ting Yao, and Tao Mei. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*.
- [28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.
- [29] Sachin Ravi and Hugo Larochelle. 2016. Optimization as a model for few-shot learning. (2016).
- [30] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. Meta-learning with memory-augmented neural networks. In *ICML*.

- [31] Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. 2021. Only time can tell: Discovering temporal data for temporal modeling. In WACV.
- [32] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *NeurIPS* (2017).
- [33] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012).
- [34] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In CVPR.
- [35] Hao Tang, Zechao Li, Zhimao Peng, and Jinhui Tang. 2020. Blockmix: meta regularization and self-calibrated inference for metric-based meta-learning. In ACM Multimedia.
- [36] Hao Tang, Chengcheng Yuan, Zechao Li, and Jinhui Tang. 2022. Learning attention-guided pyramidal features for few-shot fine-grained recognition. *Pat*tern Recognition (2022).
- [37] Anirudh Thatipelli, Sanath Narayan, Salman Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Bernard Ghanem. 2022. Spatio-temporal relation modeling for few-shot action recognition. In CVPR.
- [38] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *NeurIPS* (2021).
- [39] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *ICML*.
- [40] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In ICCV.
- [41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS* (2017).
- [42] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *NeurIPS* (2016).
 [43] Tu Vu, Minh-Thang Luong, Quoc V Le, Grady Simon, and Mohit Iyyer. 2021.
- [43] Tu Vu, Minh-Thang Luong, Quoc V Le, Grady Simon, and Mohit Iyyer. 2021. Strata: Self-training with task augmentation for better few-shot learning. arXiv preprint arXiv:2109.06270 (2021).
- [44] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Mingqian Tang, Zhengrong Zuo, Changxin Gao, Rong Jin, and Nong Sang. 2022. Hybrid relation guided set matching for few-shot action recognition. In CVPR.
- [45] Jiamin Wu, Tianzhu Zhang, Zhe Zhang, Feng Wu, and Yongdong Zhang. 2022. Motion-modulated temporal fragment alignment network for few-shot action recognition. In CVPR.
- [46] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip HS Torr, and Piotr Koniusz. 2020. Few-shot action recognition with permutation-invariant attention. In ECCV.
- [47] Ji Zhang, Lianli Gao, Xu Luo, Hengtao Shen, and Jingkuan Song. 2023. DETA: Denoised Task Adaptation for Few-Shot Learning. arXiv preprint arXiv:2303.06315 (2023).
- [48] Ji Zhang, Jingkuan Song, Lianli Gao, Ye Liu, and Heng Tao Shen. 2022. Progressive meta-learning with curriculum. *TCSVT* (2022).
- [49] Xueting Zhang, Debin Meng, Henry Gouk, and Timothy M Hospedales. 2021. Shallow bayesian meta learning for real-world few-shot recognition. In *ICCV*.
- [50] Yizhou Zhao, Xun Guo, and Yan Lu. 2022. Semantic-aligned Fusion Transformer for One-shot Object Detection. In CVPR.
- [51] Andrey Zhmoginov, Mark Sandler, and Maksym Vladymyrov. 2022. Hypertransformer: Model generation for supervised and semi-supervised few-shot learning. In *ICML*.
- [52] Linchao Zhu and Yi Yang. 2018. Compound memory networks for few-shot video classification. In ECCV.
- [53] Linhai Zhuo, Yuqian Fu, Jingjing Chen, Yixin Cao, and Yu-Gang Jiang. 2022. Tgdm: Target guided dynamic mixup for cross-domain few-shot learning. In ACM Multimedia.