Alignment-guided Temporal Attention for Video Action Recognition

Anonymous Author(s)
Affiliation
Address
email

Abstract

Temporal modeling is crucial for various video learning tasks. Most recent approaches employ either factorized (2D+1D) or joint (3D) spatial-temporal operations to extract temporal contexts from the input frames. While the former is more efficient in computation, the latter often obtains better performance. In this paper, we attribute this to a dilemma between the sufficiency and the efficiency of interactions among various positions in different frames. These interactions affect the extraction of task-relevant information shared among frames. To resolve this issue, we prove that frame-by-frame alignments have the potential to increase the mutual information between frame representations, thereby including more task-relevant information to boost the effectiveness. Then we propose Alignment-guided Temporal Attention (ATA) to extend 1-dimensional temporal attention with parameter-free patch-level alignments between neighboring frames. It can act as a general plug-in for image backbones to conduct the action recognition task without any model-specific design. Extensive experiments on multiple benchmarks demonstrate the superiority and generality of our module.

1 Introduction

Unlike images, videos include a wealth of temporal information due to common and distinct patterns in nearby frames. In images, the whole scene is static. In videos, however, the background, the foreground, and each part of them can have different degrees of movement. These differences necessitate temporal modeling in video learning tasks.

With respect to this, recent literature mainly focuses on two branches of modelings, namely, factorized (2D+1D) spatial-temporal operations [56, 15, 3, 36] and joint (3D) spatial-temporal operations [46, 11, 39, 32]. The former considers staggering or dividing spatial operations with temporal ones, such as separable 3D convolutions and factorized spatial-temporal attention. Since the two groups of operations are relatively independent, this paradigm leverages indirect interactions for features at different space-time locations. Enjoying the efficiency from fewer participants for each operation, it often suffers from less sufficiency of cross-frame cross-location interactions. In contrast, the latter extends 2-dimensional operations to their 3-dimensional varieties, thereby jointly interacting features from neighboring areas as well as frames, e.g., 2D convolutions vs. 3D convolutions and spatial attention vs. joint spatial-temporal attention. This line of work explores a more global view compared with the former one, bringing more direct and adequate interactions at the cost of dramatically increased complexity. As a result, a conflict seemingly exists between the sufficiency and the efficiency of interactions among various positions in different frames.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

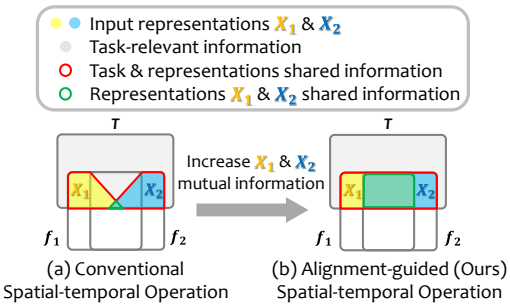

Figure 1: **Information diagram illustrating our motivation.** T stands for task-relevant information, and X_1, X_2 denote representations of adjacent input frames f_1, f_2.

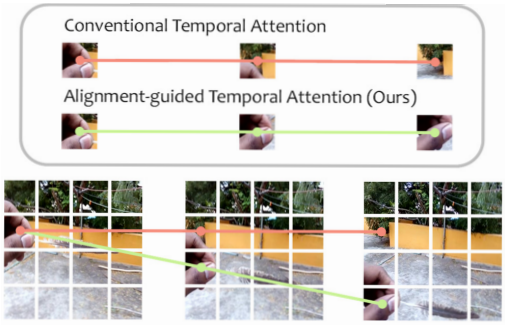

Figure 2: **Comparison between conventional Temporal Attention and Alignment-guided Temporal Attention (ATA).** Participants of each operation are connected with lines.

To address this conflict, we revisit both schemes from the perspective of information theory. With the help of more direct interactions, joint spatial-temporal operations share more mutual information between neighboring frame representations than their factorized counterparts. This potentially increases task-relevant information extracted by the model, benefiting its performance. Nevertheless, for joint spatial-temporal modeling, the ability to exploit mutual information from consecutive frames depends on the kernel size of convolution-based operations or the window size of attention-based ones. To reach the theoretical upper bound of mutual information shared, completely global interactions are needed for this paradigm, which is not realistic. In general cases, these conventional schemes collect inadequate task-relevant information as illustrated in Fig. 1(a).

Alternatively, we propose to directly interact between the most similar parts of adjacent frames, instead of among regions at nearby spatial locations. In other words, we guide the spatial-temporal operation with alignments as shown in Fig. 2. We theoretically prove that this guidance increases mutual information of neighboring frame representations as in Fig. 1(b) and practically implement this concept as our *Alignment-guided Temporal Attention* (ATA). Upon traditional Temporal Attention, it has no extra parameters and low additional computation cost. For alignment, we match patches of adjacent frames with the Kuhn-Munkres Algorithm (KMA) [4] based on cosine similarity. After temporal attention through the aligned route, we de-align patches to keep their original spatial order.

Our contributions can be summed up as follows.

1. We discuss the problems of factorized (2D+1D) and joint (3D) spatial-temporal operations and propose Alignment-guided Temporal Attention (ATA) for temporal modeling. Taking advantages of both, ATA achieves strong performance while maintaining low computational overheads.
2. We theoretically prove that the frame-by-frame alignment increases mutual information between neighboring frame representations, thereby potentially including more task-relevant information.
3. Through qualitative and quantitative experiments, we demonstrate that our ATA is superior to conventional temporal modeling methods by providing more effective cross-frame cross-location interactions. Implementations on video frameworks and image backbones (TimeSformer [3], CycleMLP [6], and ConvNeXt [33]) show its potential as a general plug-and-play module to bridge from image recognition to video action recognition.

2 Related Work

Action recognition. Early action recognition research [27, 49, 48, 47, 40, 37, 53, 13] focuses on hand-crafted features. Deep network models, on the other hand, gradually surpass traditional methods based on DT [49] and iDT [48], resulting them to disappear. To be more specific, AlexNet [24] generates numerous important video backbones, which include 2D/3D CNN [22, 1, 29, 34] in two-stream models and its variants with LSTM/GRU/RNNs [9, 42, 57]. Video backbones such as VTN [35], TimeSformer [3], ViViT [2], and Video Swin [32] have recently emerged as a result of the Transformers-based ViT [10].

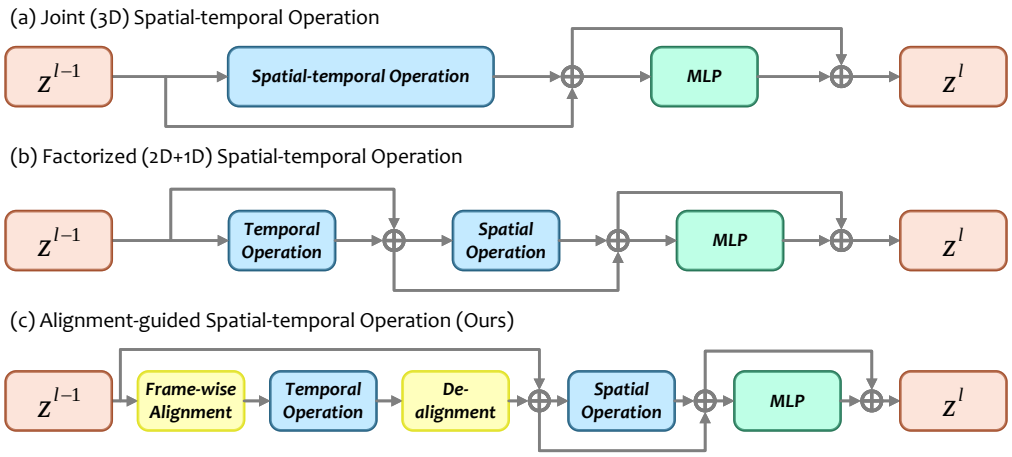

Figure 3: **Comparison among spatial-temporal operations in terms of model structures.**

Visual learning architectures. Due to the obvious ViT [10]'s high generalization, numerous ViT-based image backbones recently expand their work range into video. TimeSformer [3] presents four kinds of factorized space-time implements utilizing the standard ViT's structure. VTN [35] uses a temporal attention model to aggregate information from all video clips. MViT [11] creates a multi-scale architecture through integrating pooling and downscale methods, which have both been used in image backbones. ViViT [2] creates three variants that focus on factorized space-time attention. It leverages tubelet embedding to extend the image patches into the temporal dimension. MaskFeat's [52] video model is based on the unsupervised image work MAE [16], which applies three masking strategies: frames, cubes, and tubes.

Temporal correspondences. In comparison to image recognition, motion representation is the most difficult aspect of video action recognition [26, 21, 38, 19, 36, 41, 54, 55]. Recent alignment methods have primarily focused on supervised temporal correspondences. For representing temporal relationships, optical flow [17, 18, 43–45] is proposed to establish the displacements in pixels between successive frames. Patch-level interconnections can reduce memory cost of optical flow. For instance, [30] presents a typical descriptor model called SIFT flow and the SIFT features are being used for matching patches. Local features such as HOG [25, 7] and MBH [8] can be used to describe statistical features as well. The correlation of patches or other inter-frame components, in particular, has positive effects for models learning visual representations, according to [45].

3 Method

For video learning tasks, we consider their inputs a combination of spatial and temporal information. The former is introduced by sequential static frames, while the latter originates from their potential connections. These two parts are usually tackled jointly in convolution-based methods [5] as well as in attention-based approaches [31], with their per-block model structures roughly demonstrated in Fig. 3(a), i.e., joint (3D) spatial-temporal operations. We also notice attempts [56, 2, 3] separating spatial operations from temporal ones as in Fig. 3(b), known as factorized (2D+1D) spatial-temporal operations. For either scheme, the ability to extract mutual information from consecutive frames is restricted by the kernel/window size of its temporal operation. To mitigate this, we propose to conduct frame-wise alignment before each temporal operation and de-alignment after that as in Fig. 3(c).

3.1 Problem definition

To begin with, we define alignment generally. Given two neighboring frames $X^{t-1}, X^t \in \mathbb{R}^{HW \times C}$ with their patch embeddings $\{x_i^{t-1}\}_{i=1}^{HW}, \{x_j^t\}_{j=1}^{HW}$, the alignment function of X^{t-1} and X^t is

$$\alpha\left(X^{t-1}, X^t\right) = \arg\max_{\rho} \left\{\chi\left[X^{t-1}, \rho(X^t)\right]\right\} \tag{1}$$

where χ is a criterion function and $\rho : \mathbb{R}^{HW\times C} \rightarrow \mathbb{R}^{HW\times C}$ is a projection. The de-alignment function α^{-1} is defined as the reverse process of alignment

$$\alpha^{-1}\left(\alpha\left(X^{t-1}, X^{t}\right)\right) = X^{t} \tag{2}$$

Specifically, we can have spatial rearrangement as a series of projection ρ

$$\rho^{\text{Rearrange}}(X^{t}) = R \times X^{t} = \hat{X}^{t} = \left\{\hat{x}_{j}^{t}\right\}_{j=1}^{HW} \tag{3}$$

where the rearrangement matrix

$$\begin{aligned} R &= \left\{r_{i,j}^{t-1,t}\right\} \in \{0,1\}^{HW\times HW} \\ \text{s.t.} &\sum_{i=1}^{HW} r_{i,j}^{t-1,t} = 1, \sum_{j=1}^{HW} r_{i,j}^{t-1,t} = 1 \end{aligned} \tag{4}$$

and cosine similarity as the criterion χ

$$\chi^{\text{Cosine}}\left(X^{t-1}, \hat{X}^{t}\right) = \sum_{i=j} \frac{x_{i}^{t-1} \cdot \hat{x}_{j}^{t}}{\left\|x_{i}^{t-1}\right\| \left\|\hat{x}_{j}^{t}\right\|} \tag{5}$$

In this case, the rearrangement can be regarded as a patch-wise match between X^{t-1} and X^{t}, and the alignment should be a perfect matching scheme. As a result, the alignment matrix is one of the rearrangement matrices satisfying

$$A = \arg\max_{R}\left[\chi^{\text{Cosine}}\left(X^{t-1}, R \times X^{t}\right)\right] \tag{6}$$

and the de-alignment matrix is simply its transpose

$$D = A^{\mathrm{T}} \tag{7}$$

3.2 Proof and analysis

Mutual information of two random variables measures the information shared between them. It represents the degree of reduction in uncertainty of one variable given the other. Traditional 1D temporal operations interact features of consecutive frames at a certain spatial position. Since these features might diverse greatly, the irrelevances of them can lead to uncertainty, decreasing mutual information collected by the model. We will prove that the mutual information of adjacent frames increases after alignment, which also indicates that the similarity of adjacent frames rises.

We define the representations of two adjacent frames as random variables X^{t-1} and X^{t}. Within the same video clip, they obey the same distribution. Thus their mutual information can be calculated as

$$\text{MI}(X^{t-1}, X^{t}) = H(X^{t}) - H(X^{t} \mid X^{t-1}) \tag{8}$$

Eq. (8) expresses the mutual information between the feature X^{t-1} of the $(t-1)^{\text{th}}$ frame and the feature X^{t} of the tth frame. The definitions of information entropy $H(X^{t})$ $(t \in \{1, T\})$ and conditional entropy $H(X^{t}|X^{t-1})$ are as follows.

$$H(X^{t}) = -\sum_{j=1}^{HW} p\left(x_{j}^{t}\right) \log p\left(x_{j}^{t}\right) \tag{9}$$

$$\begin{aligned} H(X^{t} \mid X^{t-1}) &= \sum_{x_{i}^{t-1}} p(x_{i}^{t-1}) H(X^{t} \mid X^{t-1} = x_{i}^{t-1}) \\ &= -\sum_{x_{i}^{t-1}} p(x_{i}^{t-1}) \sum_{x_{j}^{t}} p(x_{j}^{t} \mid x_{i}^{t-1}) \log(p(x_{j}^{t} \mid x_{i}^{t-1})) \end{aligned} \tag{10}$$

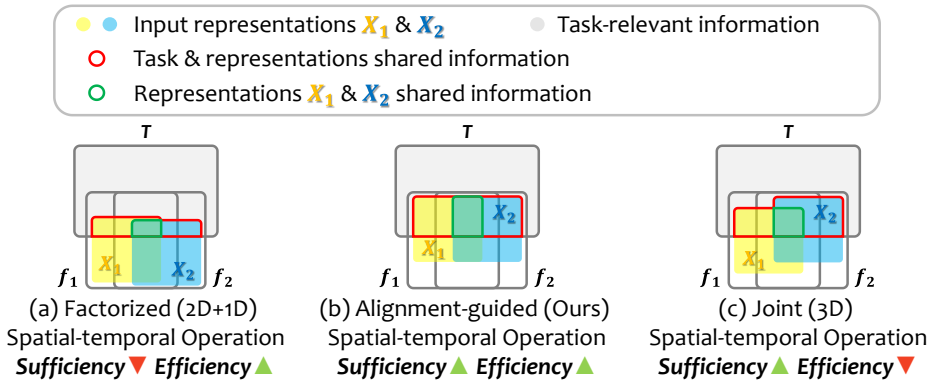

Figure 4: **Comparison of spatial-temporal operations in terms of information diagrams.** T represents task-relevant information, and X_1, X_2 are representations of adjacent input frames f_1, f_2.

where $x_i^{t-1} \in F^{t-1}$ and $x_j^t \in F^t$. F^{t-1} and F^t denote the feature sets of all patches in the $(t-1)^{\text{th}}$ frame and the tth frame, respectively. $p(x_i^{t-1})$ is the probability of occurrence of random variable X^{t-1}. $p(x_j^t \mid x_i^{t-1})$ is the conditional probability of occurrence of random variable X^t under the condition determined by random variable X^{t-1}. $p(x_j^t, x_i^{t-1})$ is the joint probability that random variables X^{t-1} and X^t satisfy certain conditions at the same time.

Given patch embeddings of a certain frame, the probability of a certain patch embedding equaling to a certain vector is determined, i.e.,

$$p\left(x_j^t\right) = p\left(\hat{x}_j^t\right) = \frac{1}{HW} \tag{11}$$

where x_j^t indicates the feature of the jth patch in the tth frame before the alignment, and $\hat{x}_j^t$ means the feature of the jth patch in the tth frame after the alignment.

Hence Eq. (10) can be simplified to

$$H(X^t \mid X^{t-1}) = -\frac{1}{HW} \sum_{x_i^{t-1}, x_j^t} p(x_j^t \mid x_i^{t-1}) \log(p(x_j^t \mid x_i^{t-1})) \tag{12}$$

The rearrangement-based alignment does not change the feature contents of the patch, thus the information entropy does not change after alignment. As a result, the mutual information of the features of two adjacent frames after the alignment is

$$\text{MI}(X^{t-1}, \hat{X}^t) = H(\hat{X}^t) - H(\hat{X}^t \mid X^{t-1}) = H(X^t) - H(\hat{X}^t \mid X^{t-1}) \tag{13}$$

We notice that the difference between Eq. (8) and Eq. (13) lies in the different conditional entropy $H(X^t \mid X^{t-1})$ and $H(\hat{X}^t \mid X^{t-1})$. With $H(X^t \mid X^{t-1})$ expressed in Eq. (12), $H(\hat{X}^t \mid X^{t-1})$ can be similarly written as

$$H(\hat{X}^t \mid X^{t-1}) = -\frac{1}{HW} \sum_{x_i^{t-1}, \hat{x}_j^t} p(\hat{x}_j^t \mid x_i^{t-1}) \log(p(\hat{x}_j^t \mid x_i^{t-1})) \tag{14}$$

We assume that two patches which are close to each other have a greater probability to be similar than the probability to be dissimilar. If we align the patches according to the Eq. (6), more patches with high similarity have the chance to be put in close positions. In this circumstance, patches with high similarity will have a greater probability of appearing at temporally close positions in adjacent frames. On the one hand, when the positions of i and j are close, the probability of similar features appearing in such close locations becomes higher. On the other hand, when the positions of i and j are far away, the dissimilar features are more likely to occur at such a distance too. In general, the conditional probability $p(\hat{x}_j^t \mid x_i^{t-1})$ will become greater than $p(x_j^t \mid x_i^{t-1})$. Consequently, Eq. (10) is greater than Eq. (14)

$$H(X^t \mid X^{t-1}) > H(\hat{X}^t \mid X^{t-1}) \tag{15}$$

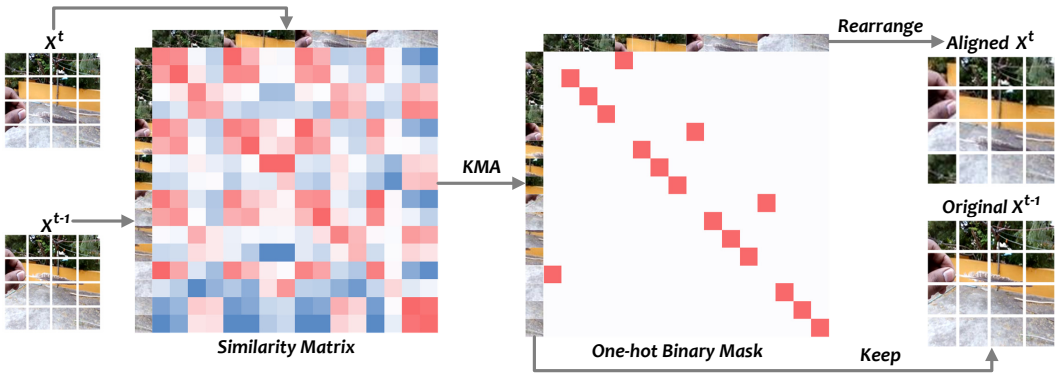

Figure 5: **Alignment with Kuhn-Munkres Algorithm (KMA).**

As the conditional entropy is negatively correlated with the mutual information, the conditional entropy decreases after alignment, and the mutual information $\mathrm{MI}(X^{t-1}, X^t)$ increases correspondingly.

$$\mathrm{MI}(X^{t-1}, \hat{X}^t) > \mathrm{MI}(X^{t-1}, X^t) \tag{16}$$

Performance of a model is related to how much task-relevant information it collects from its input [50]. In video learning tasks, this task-relevant information is a subset of frame representations, which are information extracted from input frames. For simplicity, we present a instance of two neighboring frames demonstrated in Fig. 4. Under the same model capacity, information extracted from each frame is approximately the same amount for each model. Nevertheless, the situation is different in task-relevant information. When modeling with factorized (2D+1D) spatial-temporal operations in Fig. 4(a), the shared information of frame representations is rather small due to the misalignment of patches participating in a single temporal operation. With a more global spatial-temporal view, joint (3D) operations in Fig. 4(c) tend to increase this shared information. It also results in more task-relevant information extracted but is still restricted by the kernel/window size. Since enlarging kernel/window size cubically raises the complexity, this line of work is often less efficient. As for our alignment-guided spatial-temporal operation in Fig. 4(b), it can increase the mutual information shared by neighboring frame representations with simple operations, thereby having the potential to collect more task-relevant information at a low cost. This helps with both the sufficiency and the efficiency of cross-frame cross-location interactions and leads to a better performance.

3.3 Solution

In order to solve the problem discussed, the key is to obtain an alignment matrix A. To this end, potential algorithms are supposed to produce a one-on-one matching matrix according to the given criterion. This can be simplified as a classical bipartite matching problem, which is efficient to solve through Kuhn-Munkres Algorithm (KMA) [4].

3.3.1 Alignment with Kuhn-Munkres Algorithm

Given the cosine similarity criterion as in Eq. (5), the alignment procedure with KMA is illustrated in Fig. 5. Firstly, we calculate the similarity matrix of two adjacent frames X^{t-1} and X^t

$$S^{t-1,t} = \sum_{i=j} \frac{x_i^{t-1} \cdot x_j^t}{\|x_i^{t-1}\| \|x_j^t\|} \tag{17}$$

Based on this, KMA produces a one-hot binary matrix matching patch embeddings from the previous frame and patches from the current

$$A = \mathrm{KMA}\left(S^{t-1,t}\right) \tag{18}$$

Table 1: Comparison to state-of-the-art on Kinetics-400. We report the inference cost with a single "view" (temporal clip with spatial crop), the same below.

Method	Pretrain	Frames	Top-1 Acc	Top-5 Acc	Views	FLOPs (G)	Param (M)
ViT-B-VTN [35]	IN-21K	16	79.8	94.2	1×1	4218	114
VidTr-L [28]	IN-21K	32	79.1	93.9	3×10	351	-
ViViT-L [2]	IN-21K	16	**81.3**	94.7	3×4	3992	310.8
TimeSformer [3]	IN-21K	8	78.0	93.7	3×1	196	121.4
TimeSformer-L [3]	IN-21K	96	80.7	94.7	3×1	2380	121.4
TimeSformer-HR [3]	IN-21K	16	79.7	94.4	3×1	1703	121.4
MViTv1-B [11]	K400	16	78.4	93.5	1×5	70.5	36.6
MViTv1-B [11]	K400	32	80.2	94.4	1×5	170	36.6
MViTv1-B [11]	K400	64	81.2	95.1	3×3	455	36.6
Mformer-B [36]	IN-21K	16	79.7	94.2	3×10	369.5	109.1
Mformer-L [36]	IN-21K	32	80.2	94.8	3×10	1185.1	109.1
Mformer-HR [36]	IN-21K	16	81.1	**95.2**	3×10	958.8	381.9
ATA (Ours)	IN-21K	32	**81.4**	**95.5**	3×1	792.9	121.8

Then the spatial order of the previous frame features is kept, and that of the current is rearranged according to Eq. (3). This process is applied to every neighboring frame pairs, thereby producing an aligned route for 1D temporal operations. Since the computational complexity of KMA is $O(N^3)$ [20], alignment with KMA have a complexity of $O(TH^3W^3)$.

3.3.2 De-alignment

After aligning frame representations with rearrangement, the spatial structures of them are modified. To keep the original location-based information and facilitate subsequent spatial operations, a recovery with de-alignment as in Eq. (7) is thus needed. As a result, the alignment, the temporal operation, and the de-alignment as a whole form a temporal interaction through a 3D path, within which each point on the path is similar to the next one.

4 Experiments

Datasets. We employ two widely used video action recognition datasets, i.e., Kinetics-400 (K400) [23] and Something-Something V2 (SSv2) [14], in our experiments. Kinetics-400 contains 240k training videos and 30k validation videos in 400 classes of human actions. Something-Something V2 consists of 168.9K training videos and 24.7K validation videos for 174 classes. We provide the the top-1 and top-5 accuracy on the validation sets, as well as the inference complexity measured with FLOPS.

Training and inference. We employ TimeSformer [3] as our baseline, and finetune on video datasets using the officially released model pretrained on ImageNet-21K [24]. Standard augmentations and a head dropout rate of 0.5 as in [12] are also employed. We adopt SGD to optimize our network for 30 epochs with a batch size of 64. The initial learning rate is set to 0.005 with $0.1\times$ decays on the 21st and 27th epochs. All patch embeddings adopt a weight decay of 1e-4. No weight decay is used for the class token and the positional embedding. In ablation study, we provide the results of 8-frame input on Kinetics-400. And in the main experiments compared with the state-of-the-art methods, we provide the results of 32-frame input. The resolution of 224×224 is used throughout the experiments.

4.1 Comparison to state-of-the-art

Kinetics-400. Table 1 presents comparison to state-of-the-art results on Kinetics-400. As our motivation is to demonstrate the effectiveness of alignment without bells and whistles, we compare our model only with the approaches based on ViT to reduce the impact from architecture inconsistency.

Table 2: Comparison to state-of-the-art on Something-Something V2.

Method	Pretrain	Frames	Top-1 Acc	Top-5 Acc	Views	FLOPs (G)	Param (M)
ViViT-L [2]	IN-21K+K-400	16	65.4	89.8	-	903	352.1
TimeSformer [3]	IN-21K	8	59.5	74.9	3×1	196.7	121.4
TimeSformer-L [3]	IN-21K	96	62.4	81.0	3×1	2380	121.4
TimeSformer-HR [3]	IN-21K	16	62.2	78.0	3×1	1703	121.4
MViTv1-B [11]	K-400	16	64.7	89.2	3×1	70.5	36.6
MViTv1-B [11]	K-400	32	67.1	90.8	3×1	170	36.6
MViTv1-B [11]	K-400	64	**67.7**	90.9	3×1	455	36.6
Mformer-B [36]	IN-21K+K-400	16	66.5	90.1	3×1	369.5	109.1
Mformer-L [36]	IN-21K+K-400	32	**68.1**	**91.2**	3×1	1185.1	109.1
Mformer-HR [36]	IN-21K+K-400	16	67.1	90.6	3×1	958.8	381.9
ATA (Ours)	IN-21K	32	67.0	**91.0**	3×1	792.9	121.8
ATA (Ours)	IN-21K	32	67.1	90.8	4×3	792.9	121.8

All these methods are pretrained on ImageNet-21K except for MViT [11], which is trained from scratch. Results show that our method outperforms most of existing ViT-based approaches. It is noteworthy that our method has the same factorized spatial-temporal structure but different form of temporal modeling with TimeSformer [3]. Using less input frames and lower resolution, ATA even surpasses TimeSformer-L and TimeSfomer-HR, which can fully demonstrates its effectiveness. Compared with Motionformer [36], which explores motion trajectory in temporal attention, our method achieves better performance with much lower complexity. Our approach can achieve better performance if using pretrained models on larger datasets, e.g., JFT-300M.

Something-Something v2. Table 2 compares our approach with the state-of-the-art methods on SSv2. Different from Kinetics-400, SSv2 consists of videos with high temporal reasoning. Therefore, pretraining on video data would be very helpful. That may be the reason why our method shows lower performance than Motionformer-L at the same input setting. It is noteworthy that our method achieves comparable result to MViT-B with 32-frame input. Considering its 3D structure and pretraining on Kinetics-400, our method demonstrates stronger ability in temporal modeling.

4.2 Ablation study

To better understand our approach, we ablate our ATA on the Kinetics-400 dataset with 8-frame inputs. We prove its effectiveness by adding it to different locations of our baseline model. Also, we demonstrate its generality by comparing it with different temporal modeling methods on Convolution/MLP/Attention-based image backbones.

ATA module in different locations. We divide 12 blocks of our baseline into 4 stages, each containing 3 blocks. Then we insert our ATA progressively, as shown in Table 3. Generally, the models perform better with ATA compared with that without it. The presence of ATA clearly promotes temporal attention wherever it appears, supporting the conclusion that adding mutual information can improve performance. Recognition accuracy grows gradually with ATA added to more stages, with a maximum growth of 1.6%. The model benefits partially from aligning features at different depths, thus aligning features from disparate hyperspaces and increasing mutual information for every block of features. We may even further reduce the compute time and FLOPS from ATA's utilization in a new backbone. It is also worthwhile to notice that when ATA is added to only one single stage, the model gains 1.3% compared with plain temporal attention. This indicates that alignment is effective even if only added to a small range of temporal operations.

ATA module on different image backbones. To investigate the generality of ATA, we add three kinds of temporal modeling to CycleMLP [6], ConvNeXt [33] and ViT-Base [10]. Specifically, averaging means computing the mean of all spatial-temporal features, attention refers to factorized 1D temporal attention, and ATA is our proposed module. By using parametric attention, the image backbones obtain significant gains of 1.9%, 0.9% and 2.0% accordingly compared with averaging. Furthermore,

Table 3: Ablation study of ATA module in different locations.

Block 0-2	Block 3-5	Block 6-8	Block 9-11	Top-1 Acc	Top-5 Acc	FLOPs (G)	Param (M)
-	-	-	-	78.0	93.7	201.8	121.8
✓	-	-	-	79.3	94.4	201.8	121.8
✓	✓	-	-	79.2	94.3	201.8	121.8
✓	✓	✓	-	79.4	94.4	201.8	121.8
✓	✓	✓	✓	79.6	94.3	201.8	121.8

Table 4: Ablation study of ATA module on different image backbones.

Image Backbone	Temporal Modeling	Top-1 Acc	Top-5 Acc	FLOPs (G)	Param (M)
CycleMLP-B5 [6]	Averaging [51]	74.9	92.1	75.3	82.4
CycleMLP-B5 [6]	Attention [3]	76.8	93.1	122.4	102.8
CycleMLP-B5 [6]	ATA (Ours)	77.7	93.5	122.4	102.8
ConvNeXt-Base [33]	Averaging [51]	79.0	94.0	123.0	88.0
ConvNeXt-Base [33]	Attention [3]	80.1	94.8	198.0	140.5
ConvNeXt-Base [33]	ATA (Ours)	80.5	94.8	198.0	140.5
ViT-Base [10]	Averaging [51]	76.0	92.6	140.3	86.2
ViT-Base [10]	Attention [3]	78.0	93.7	201.8	121.8
ViT-Base [10]	ATA (Ours)	79.6	94.3	201.8	121.8

our ATA achieves extra scores of 0.9%, 0.6% and 1.6% on the basis of temporal attention. Temporal attention aggregate temporal features before calculating the mean of the features. Attention will weight and sum other frames to the present frame, so that each frame has the information of all other frames, and their mutual information will increase in the temporal. As the mutual information improves, the performance of the backbones with attention are better than those with ordinary averaging. What's more, ATA rearranges the similar patches between frames on the same coordinate. We can conclude from Section 3.2 that our ATA increases the mutual information on the basis of temporal attention, so the backbones with ATA can perform better than those with temporal attention. As a result, our ATA helps them all perform better without extra FLOPs and parameters.

5 Conclusion

In this paper, we investigate temporal modeling approaches from the perspective of the sufficiency and the efficiency of spatial-temporal interactions. Through theoretical analysis, we find that increasing mutual information in neighboring frame representations can potentially promote task-relevant information collected by the model, and thus benefits its performance. Then we propose a general alignment and de-alignment scheme to enlarge the mutual information among consecutive frames. For specific implementation, we propose our Alignment-guided Temporal Attention (ATA) to conduct feature-level alignment and similarity-based temporal modeling. Extensive experiments show the effectiveness and generality of ATA. Our proposed module not only achieves satisfying results on multiple video action recognition benchmarks, but also can act as a general plug-in to extend any image backbone for video learning tasks.

Future work. We will improve our ATA in terms of parameterization. For instance, the process of matching features from adjacent frames can be potentially conducted by a network, or this process can be modeled by consecutive displacements rather than discrete reordering of features. In these cases, learnable networks and displacements enable end-to-end training of the model, leading to a higher upper bound of performance.

Potential negative societal impacts. From the perspective of applications, we consider the abuse of our video action recognition models. For example, unauthorized surveillance from hidden monitors will violate the privacy rights of people, especially in hotels and homes. To alleviate such ethical problems, the permission of using the model should be restricted.

References

[1] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *CVPR*, pages 5297–5307, 2016.

[2] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid. Vivit: A video vision transformer. In *ICCV*, pages 6836–6846, 2021.

[3] G. Bertasius, H. Wang, and L. Torresani. Is space-time attention all you need for video understanding. *arXiv preprint arXiv:2102.05095*, 2(3):4, 2021.

[4] D. P. Bertsekas. A new algorithm for the assignment problem. *Mathematical Programming*, 21 (1):152–171, 1981.

[5] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017.

[6] S. Chen, E. Xie, C. Ge, D. Liang, and P. Luo. Cyclemlp: A mlp-like architecture for dense prediction. *arXiv preprint arXiv:2107.10224*, 2021.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893. Ieee, 2005.

[8] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, pages 428–441. Springer, 2006.

[9] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634, 2015.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021.

[12] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019.

[13] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars. Rank pooling for action recognition. *PAMI*, 39(4):773–787, 2016.

[14] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, pages 5842–5850, 2017.

[15] X. Guo, X. Guo, and Y. Lu. Ssan: Separable self-attention network for video representation learning. In *CVPR*, pages 12618–12627, 2021.

[16] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

[17] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3): 185–203, 1981.

[18] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, pages 2462–2470, 2017.

[19] A. Jabri, A. Owens, and A. Efros. Space-time correspondence as a contrastive random walk. *NeurIPS*, 33:19545–19560, 2020.

[20] T. Jebara. Kernelizing sorting, permutation, and alignment for minimum volume pca. In *International Conference on Computational Learning Theory*, pages 609–623. Springer, 2004.

[21] S. Jiang, D. Campbell, Y. Lu, H. Li, and R. Hartley. Learning to estimate hidden motions with global motion aggregation. In *ICCV*, pages 9772–9781, 2021.

[22] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732, 2014.

[23] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 25, 2012.

[25] A. Kushwaha and A. Khare. Human activity recognition by utilizing local ternary pattern and histogram of oriented gradients. In *ICBMA*, pages 315–324. Springer, 2021.

[26] H. Kwon, M. Kim, S. Kwak, and M. Cho. Motionsqueeze: Neural motion feature learning for video understanding. In *ECCV*, pages 345–362. Springer, 2020.

[27] I. Laptev. On space-time interest points. *IJCV*, 64(2):107–123, 2005.

[28] X. Li, Y. Zhang, C. Liu, B. Shuai, Y. Zhu, B. Brattoli, H. Chen, I. Marsic, and J. Tighe. Vidtr: Video transformer without convolutions. *arXiv e-prints*, pages arXiv–2104, 2021.

[29] R. Lin, J. Xiao, and J. Fan. Nextvlad: An efficient neural network to aggregate frame-level features for large-scale video classification. In *ECCV Workshops*, pages 0–0, 2018.

[30] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *PAMI*, 33:978–94, 05 2011. doi: 10.1109/TPAMI.2010.147.

[31] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021.

[32] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021.

[33] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022.

[34] A. Miech, I. Laptev, and J. Sivic. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*, 2017.

[35] D. Neimark, O. Bar, M. Zohar, and D. Asselmann. Video transformer network. In *ICCV*, pages 3163–3172, 2021.

[36] M. Patrick, D. Campbell, Y. Asano, I. Misra, F. Metze, C. Feichtenhofer, A. Vedaldi, and J. F. Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *NeurIPS*, 34, 2021.

[37] J. S. Pérez, E. Meinhardt-Llopis, and G. Facciolo. Tv-l1 optical flow estimation. *IPOL*, 2013: 137–150, 2013.

[38] D. Ramanan, D. A. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *CVPR*, volume 1, pages 271–278. IEEE, 2005.

[39] M. S. Ryoo, A. Piergiovanni, A. Arnab, M. Dehghani, and A. Angelova. Tokenlearner: Adaptive space-time tokenization for videos. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=z-l1kpDXs88`.

[40] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 105(3):222–245, 2013.

[41] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain. Time-contrastive networks: Self-supervised learning from video. In *ICRA*, pages 1134–1141. IEEE, 2018.

[42] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *NeurIPS*, 28, 2015.

[43] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *CVPR*, pages 2432–2439. IEEE, 2010.

[44] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, pages 8934–8943, 2018.

[45] Z. Teed and J. Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419. Springer, 2020.

[46] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015.

[47] M. Vochten, T. De Laet, and J. De Schutter. Generalizing demonstrated motion trajectories using coordinate-free shape descriptors. *Robotics and Autonomous Systems*, 122:103291, 2019.

[48] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, pages 3551–3558, 2013.

[49] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 103(1):60–79, 2013.

[50] H. Wang, X. Guo, Z.-H. Deng, and Y. Lu. Rethinking minimal sufficient representation in contrastive learning. *arXiv preprint arXiv:2203.07004*, 2022.

[51] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36. Springer, 2016.

[52] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. *arXiv preprint arXiv:2112.09133*, 2021.

[53] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. In *ICCV*, pages 1385–1392, 2013.

[54] J. Wills, S. Agarwal, and S. Belongie. What went where [motion segmentation]. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I, 2003. doi: 10.1109/CVPR.2003.1211335.

[55] M. Wu, B. Jiang, D. Luo, J. Yan, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and X. Yang. Learning comprehensive motion representation for action recognition. *arXiv preprint arXiv:2103.12278*, 2021.

[56] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, pages 305–321, 2018.

[57] J. Zhou, Y. Cao, X. Wang, P. Li, and W. Xu. Deep recurrent models with fast-forward connections for neural machine translation. *TACL*, 4:371–383, 2016.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cited all video datasets, codes, and models used in this paper.
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]