Neural Networks 195 (2026) 108313

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Full Length Article

RepAttn3D: Re-parameterizing 3D attention with spatiotemporal augmentation for video understanding

Xiusheng Lu [a], Lechao Cheng [b], Sicheng Zhao [c,*], Ying Zheng [d], Yongheng Wang [e], Guiguang Ding [a,c,*], Mingli Song [f]

[a] School of Software, Tsinghua University, Beijing, 100084, China
[b] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, 230009, China
[c] BNRist, Tsinghua University, Beijing, 100084, China
[d] Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong, China
[e] Research Center for Astronomical Computing, Zhejiang Lab, Hangzhou, 311100, China
[f] College of Computer Science, Zhejiang University, Hangzhou, 310027, China

ARTICLE INFO

Keywords:
Action recognition
Re-parameterization
Spatiotemporal coherence prior
3D Attention

ABSTRACT

The technique of structural re-parameterization has been widely adopted in Convolutional Neural Networks (CNNs) and Multi-Layer Perceptrons (MLPs) for image-related tasks. However, its integration with attention mechanisms in the video domain remains relatively unexplored. Moreover, video analysis tasks continue to face challenges due to high computational costs, particularly during inference. In this paper, we investigate the re-parameterization of widely-used 3D attention mechanism for video understanding by incorporating a spatiotemporal coherence prior. This approach allows the learning of more robust video features while introducing negligible computational overhead at inference time. Specifically, we propose a SpatioTemporally Augmented 3D Attention (STA-3DA) module as a building block for Transformer architectures. The STA-3DA integrates 3D, spatial, and temporal attention branches during training, serving as an effective replacement for standard 3D attention in existing Transformer models and leading to improved performance. During testing, the different branches are merged into a single 3D attention operation via learned fusion weights, resulting in minimal additional computational cost. Experimental results demonstrate that the proposed method achieves competitive video understanding performance on benchmark datasets such as Kinetics-400 and Something-Something V2.

1. Introduction

Video understanding plays a crucial role in a variety of applications, such as robotics (Voronin et al., 2021), sports (Wu et al., 2022), and human-computer interaction (Haria et al., 2017). Despite the emergence of numerous CNN-based approaches (Carreira & Zisserman, 2017; Gao et al., 2023, 2025; Lin et al., 2019; Tran et al., 2015), their effectiveness remains limited by the inherent lack of global modeling capabilities. Inspired by the remarkable success of Transformers in the field of Natural Language Processing, researchers have introduced various visual Transformers (Dosovitskiy et al., 2020; Fan et al., 2021; Li et al., 2023b; Liu et al., 2022; Zhou et al., 2025) for proficiently processing image and video data.

In contrast to the 2D attention commonly used in image Transformers, many recent video Transformers adopt 3D attention or its variants for joint spatiotemporal modeling. For instance, Uniformer (Li et al., 2023a) applies 3D attention in deeper layers to capture long-range token dependencies from a global perspective. Similarly, Video Swin (Liu et al., 2022) incorporates a locality inductive bias by performing self-attention within each 3D shifted window. Despite their effectiveness, these methods are often hindered by the high computational complexity of 3D attention, which limits their practical applicability. Furthermore, their deployment on edge devices presents a significant challenge, as such platforms typically impose stringent constraints on computational overhead and storage during inference. Thus, the enhancement of model performance without augmenting the computational burden of inference constitutes an important research objective.

Structural re-parameterization has been effectively used to decouple network design between training and inference, improving model performance while reducing computational cost during deployment. This technique has been successfully applied in various CNN- and MLP-based architectures. For instance, RepVGG (Ding et al., 2021b) extends the

* Corresponding authors.
E-mail addresses: xiusheng.lu.cs@gmail.com (X. Lu), chenglc@hfut.edu.cn (L. Cheng), schzhao@gmail.com (S. Zhao), zhengyinghit@outlook.com (Y. Zheng), wangyh@zhejianglab.com (Y. Wang), dinggg@tsinghua.edu.cn (G. Ding), brooksong@zju.edu.cn (M. Song).

https://doi.org/10.1016/j.neunet.2025.108313
Received 8 March 2025; Received in revised form 3 November 2025; Accepted 6 November 2025
Available online 11 November 2025

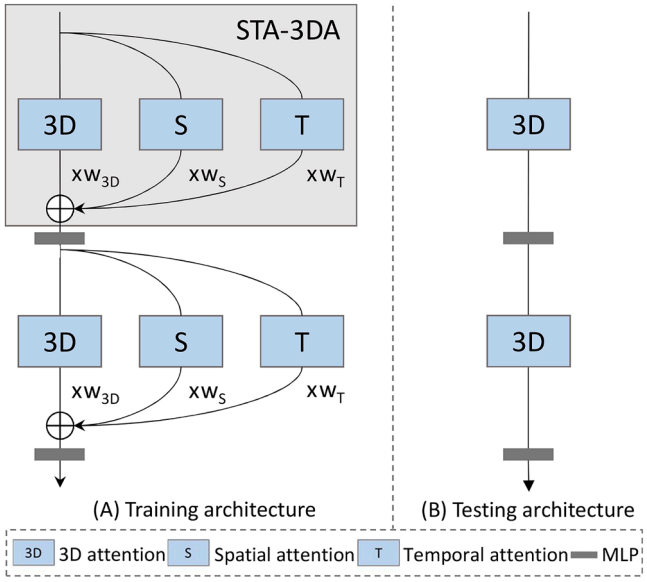

Fig. 1. An Illustration of the STA-3DA Transformer. The figure presents a partial overview of the STA-3DA Transformer built upon a ViT-B backbone. Two consecutive layers are depicted, each consisting of an STA-3DA module followed by an MLP layer. The proposed STA-3DA module enhances 3D attention by incorporating learnable spatial and temporal attention branches during training (A), which dynamically reinforce spatiotemporal relevant dependencies. During inference (B), these branches are fused, allowing the module to simplify into standard 3D attention. The rectangular blocks labeled 3D, S, and T correspond to the 3D, spatial, and temporal attention modules, respectively. The terms w_{3D}, w_S, and w_T denote the learnable weights associated with each branch.

original 3×3 kernel with identity and 1×1 branches during training, which are later merged for efficient inference. RepMLP (Ding et al., 2021a) utilizes convolutional operations to integrate local priors into fully connected layers, then fuses them together at test time. While these methods have shown success in image-based tasks, their application in video understanding remains less explored. In this work, we investigate the integration of structural re-parameterization with attention mechanisms for video modeling, extending its use beyond convolutional and fully-connected architectures.

To leverage structural re-parameterization, we first introduce a spatiotemporal coherence prior to enhance 3D attention. This prior, which captures correlations among pixels across both spatial and temporal dimensions, is widely used in video modeling. Existing methods often exploit spatiotemporal coherence by decomposing video modeling into separate appearance and motion components. For instance, certain CNN-based approaches (Tran et al., 2018) factorize 3D convolution into 2D spatial and 1D temporal convolutions, while some Transformer-based techniques (Arnab et al., 2021) decouple 3D attention into spatial and temporal attention mechanisms. In contrast to these works, our approach strengthens spatiotemporal relationships within 3D attention by introducing dedicated spatial and temporal attention branches during training. These branches are later merged via structural re-parameterization, effectively reducing inference-time computational cost while maintaining modeling capacity.

Specifically, we introduce an STA-3DA module that integrates three parallel attention pathways: 3D attention, spatial attention, and temporal attention. During training, these branches are computed concurrently and fused additively using their respective learned weights, as illustrated in Fig. 1(A). This design enables adaptive weighting of spatial and temporal dependencies associated with the current token, thereby enhancing highly relevant spatiotemporal associations. During inference, the spatial and temporal attention pathways are seamlessly merged into the 3D attention branch, shown in Fig. 1(B). This integration ensures that our method introduces negligible inference-time computational overhead compared to standard 3D attention. The merging operation is feasible because the attention matrices of spatial and temporal attention can be directly extracted as submatrices of the 3D attention matrix. By leveraging the distributive property of matrix multiplication, we combine the three attention types through summation of their respective attention matrices, followed by multiplication with the value vector. As a result, the need to compute separate attention matrices for spatial and temporal attention, along with the need to apply them to the value tensor, is eliminated during testing. This design significantly enhances the practical applicability of the proposed approach in real-world deployments.

We evaluate the proposed STA-3DA Transformer against state-of-the-art methods on several public video benchmarks, including Kinetics-400 and Something-Something V2. Comprehensive ablation studies are conducted, along with visualizations of both the attention matrices and the learned video features. The main contributions of this work are summarized as follows:

- We introduce a method that leverages spatiotemporal coherence prior to dynamically enhance relevant spatial and temporal dependencies in 3D attention, leading to the learning of more expressive video features.
- We present a novel merging technique that integrates spatial and temporal attention into a unified 3D attention mechanism, effectively converting a multi-branch training architecture into a single-branch structure at inference. This design preserves the benefits of the spatiotemporal prior while introducing negligible computational overhead.
- We propose the STA-3DA module as a building block capable of replacing standard 3D attention in video Transformers. The module strengthens the representational capacity of existing models without sacrificing practical efficiency.

2. Related work

CNN-based Video Recognition. With the development of deep learning, researchers have shifted their interest from traditional hand-crafted features such as 3D SIFT (Scovanner et al., 2007) and HOG3D (Klaser et al., 2008) to CNN-based approaches such as C3D (Tran et al., 2015) and TSM (Lin et al., 2019). Certain techniques leveraged 3D convolution (Tran et al., 2015) or decomposed it into a combination of 2D convolution and 1D convolution (Qiu et al., 2017) to extract spatiotemporal information from videos. Some other works (Li et al., 2020b; Lin et al., 2019; Wang et al., 2021) aimed to design lightweight methods by incorporating dedicatedly designed temporal modeling modules on top of 2D convolution. Despite their achievements, both categories of approaches are limited by the inherent inability of convolution operations to effectively capture long-range dependencies.

Transformer-based Video Recognition. Given the success of Transformers in image classification tasks such as ViT (Dosovitskiy et al., 2020), researchers have naturally turned to utilizing them for processing video data. TimeSformer (Bertasius et al., 2021) and ViViT (Arnab et al., 2021) disentangled the spatiotemporal analysis of videos into independent appearance and motion modeling by utilizing spatial attention and temporal attention. Uniformer (Li et al., 2023a) achieved a balance between computation and performance by separately learning local and global relations in both shallow and deep layers. MViT (Fan et al., 2021) employed a multi-stage architecture, which progressively expanded the channel dimension and reduced the spatial dimension to learn a multiscale feature pyramid. Video Swin (Liu et al., 2022) designed a 3D shifted windows based attention mechanism, which introduced a locality inductive bias to reduce computational complexity. These methods adopt a range of strategies, including attention decomposition, multi-scale feature hierarchies, and locality, to optimize the computational efficiency of 3D attention while ensuring effective modeling. On the

contrary, this paper focuses on leveraging structural re-parameterization to enhance the learning capacity of 3D attention without increasing inference-time cost, thus complementing these methods.

Structural Re-parameterization. In this paper, the structural re-parameterization technique refers to the utilization of a training-time multiple-branch structure, followed by the integration of parallel temporal and spatial attention into 3D attention during inference. In previous methods, this technique was combined with the convolution and fully-connected (FC) operations. ACNet (Ding et al., 2019) employed the asymmetric convolutions to enhance the skeletons of the conventional square convolution kernels, improving the model's invariance to rotational distortions. RepVGG (Ding et al., 2021b) built upon the foundation of 3×3 convolution by incorporating the 1×1 convolution and residual connection branches to expand the model's capacity. During testing, the model structure was simplified to a standard VGG (Simonyan & Zisserman, 2014) topology for computational efficiency. RepMLP (Ding et al., 2021a) merged the locality of convolution with the globality and positional perception of FC, and could fuse convolution into FC for inference. RepViT (Wang et al., 2024a) reconstructed the pure convolutional network of MobileNetV3 by integrating the advanced architectural principles of lightweight Vision Transformers, such as the MetaFormer structure, early convolutions, and the deeper downsampling layers. TDRL (Tan et al., 2024) incorporated re-parameterization technique into vanilla Vision Transformers, boosting model performance without incurring additional inference cost by employing linear ensemble, pyramid multi-branch structure, and distribution rectification methods. These models leverage inductive biases such as locality and apply re-parameterization to CNNs, MLPs, and Transformers, but they are primarily aimed at tasks in the image domain. In contrast to them, this paper introduces spatiotemporal coherence prior and combines structural re-parameterization technique with 3D attention for video recognition.

Spatiotemporal Coherence Prior. The pixels within the same frame or spatial location in videos exhibit inherent correlations, known as the spatiotemporal coherence prior, which has been leveraged in various approaches. Traditional hand-crafted descriptors extracted features from the surrounding 3D regions of key points (Klaser et al., 2008) or the motion trajectories (Wang et al., 2013). Some CNN-based methods employed the convolution operation to learn spatial correlations within individual frames, and then conduct temporal pooling to capture dependencies across frames. Several video Transformers (Arnab et al., 2021; Bertasius et al., 2021) decomposed 3D attention into a sequence of 2D and 1D attention to simplify computations. Distinguishing ourselves from them, we integrate weighted branches of spatial and temporal attention alongside 3D attention, then merge them together for evaluation.

3. STA-3DA Transformer

In this section, we provide a comprehensive introduction to the STA-3DA Transformer. We start by presenting the spatiotemporally augmented 3D attention module. Next, we explain the inference-time branch fusion algorithm with structural re-parameterization. Finally, we describe the proposed model instantiated with the ViT-B backbone.

3.1. Spatiotemporally augmented 3D attention module

Guided by the spatiotemporal coherence prior, the proposed STA-3DA module implicitly strengthens the important associations considered in 3D attention by introducing the parallel spatial and temporal attention branches, as depicted in Fig. 1(A). The illustration of these three types of attention is shown in Fig. 2(A). For the given token highlighted by a yellow triangle, 3D attention takes into account its dependencies with all the spatiotemporal tokens represented in cyan. In contrast, spatial attention and temporal attention calculate the relations between the given token and the tokens belonging to the same frame or spatial position, denoted by the colors grayish yellow and purple respectively.

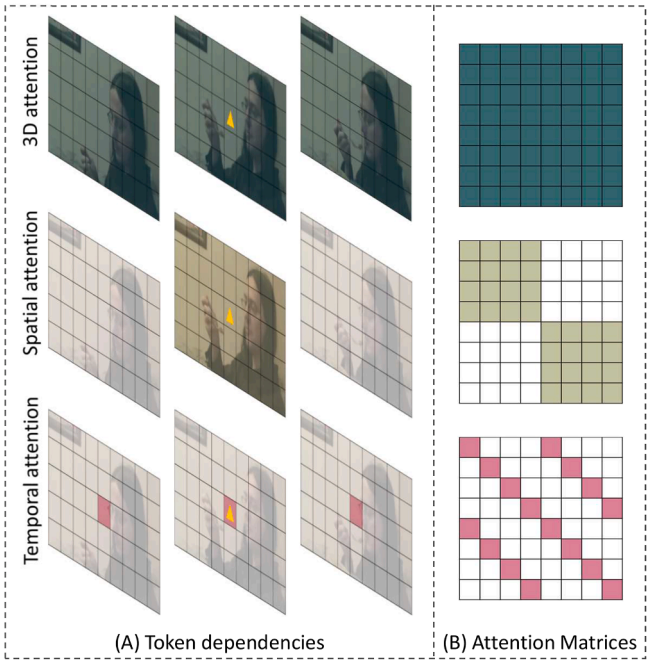

Fig. 2. Visualization of different attention mechanisms. (A) The tokens used for computing correlations with a given query token (yellow triangle) in 3D, spatial, and temporal attention mechanisms. The corresponding tokens are highlighted in cyan, grayish yellow, and purple, respectively. (B) The attention matrices, depicting the associations considered for each token across the three mechanisms.

The input tensor is symbolized as $X \in \mathbb{R}^{L\times D}$, where D represents the number of channels, and $L = L_S \times L_T$, with L_S and L_T separately indicating the spatial and temporal dimensions of the feature maps. Following Dosovitskiy et al. (2020), we first calculate the query tensor $Q_{3D} \in \mathbb{R}^{L\times D}$, key tensor $K_{3D} \in \mathbb{R}^{L\times D}$, and value tensor $V_{3D} \in \mathbb{R}^{L\times D}$ for the 3D attention through linear mappings using the embedding matrices $W_q, W_k, W_v \in \mathbb{R}^{D\times D}$. These tensors are then reshaped to derive the corresponding tensors $Q_S, K_S, V_S \in \mathbb{R}^{L_T\times L_S\times D}$ and $Q_T, K_T, V_T \in \mathbb{R}^{L_S\times L_T\times D}$ for spatial attention and temporal attention respectively

$$Q_{3D} = XW_q, K_{3D} = XW_k, V_{3D} = XW_v \tag{1}$$

$$Q_S, K_S, V_S = \mathcal{R}_S(Q_{3D}, K_{3D}, V_{3D}) \tag{2}$$

$$Q_T, K_T, V_T = \mathcal{R}_T(Q_{3D}, K_{3D}, V_{3D}) \tag{3}$$

where $\mathcal{R}_S$ and $\mathcal{R}_T$ denote the Reshape operations. For example, $\mathcal{R}_S$ can transform $Q_{3D} \in \mathbb{R}^{L\times D}$ into $Q_S \in \mathbb{R}^{L_T\times L_S\times D}$, and $\mathcal{R}_T$ into $Q_T \in \mathbb{R}^{L_S\times L_T\times D}$. These transformations enable Eq. (4) to obtain three types of correlations via matrix multiplication: global spatiotemporal, intra-frame, and same-location inter-frame token dependencies. Subsequently, the attention matrices $A_{3D} \in \mathbb{R}^{L\times L}$, $A_S \in \mathbb{R}^{L_T\times L_S\times L_S}$, and $A_T \in \mathbb{R}^{L_S\times L_T\times L_T}$ for the three attention mechanisms are computed through the following operation

$$A_{\{3D,S,T\}} = Q_{\{3D,S,T\}} K^T_{\{3D,S,T\}} / \sqrt{D} \tag{4}$$

where $\sqrt{D}$ denotes the normalization factor. Finally, the mixture of three attention can be implemented by separately applying these attention matrices to their value tensors and performing a weighted summation

$$Attn_{mix} = w_{3D} S(A_{3D}) V_{3D} + w_S S(A_S) V_S + w_T S(A_T) V_T \tag{5}$$

where w_{3D}, w_S, w_T indicate the learnable weights for the three branches, and $S(.)$ represents the Softmax function.

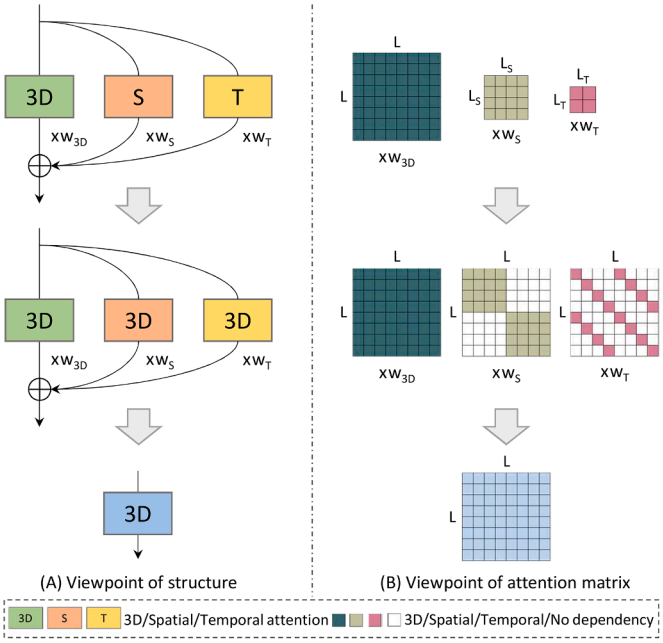

Fig. 3. Structural re-parameterization of the proposed STA-3DA module. The branch fusion from a structural perspective in (A) corresponds to the weighted sum of the expanded attention matrices shown in (B). The rectangular blocks labeled 3D, S, and T represent the 3D, spatial, and temporal attention modules, respectively. Each branch is associated with a learnable weight, denoted as w_{3D}, w_S, and w_T. The sizes of the feature maps for the 3D, spatial, and temporal attention modules are given by L, L_S, and L_T, where $L = L_S \times L_T$.

3.2. Re-param for inference-time branch fusion

During the testing phase, we try to merge the spatial and temporal attention branches to 3D attention with structural re-parameterization. Fig. 2(B) showcases the attention matrices for 3D, spatial, and temporal attention mechanisms, all adjusted to a consistent size for better visualization. Cyan, grayish yellow, and purple squares denote their respective dependencies, while white squares indicate no token dependency. It can be observed that the attention matrices of spatial and temporal attention are submatrices of that of 3D attention. For instance, the spatial attention matrix corresponds to the block diagonal region of 3D attention matrix. Hence we can directly obtain the attention matrices of spatial and temporal attention using the following steps

$$A_S = \mathcal{F}_S(A_{3D}), A_T = \mathcal{F}_T(A_{3D}) \tag{6}$$

where $\mathcal{F}_S$ and $\mathcal{F}_T$ represent the Extract operations, corresponding to the $3th$ and $4th$ steps in Algorithm 1. The role of $\mathcal{F}_S$ is to derive intra-frame attention weights from A_{3D}, whereas $\mathcal{F}_T$ is responsible for extracting inter-frame attention for spatially aligned tokens. As illustrated in Fig. 3(B), we then expand the attention matrices of spatial and temporal attention to the same size as that of 3D attention for subsequent addition fusion

$$\tilde{A}_S = \mathcal{E}_S(A_S), \tilde{A}_T = \mathcal{E}_T(A_T) \tag{7}$$

where $\mathcal{E}_S$ and $\mathcal{E}_T$ represent the Expand operations that extends the attention matrices $A_S \in \mathbb{R}^{L_T \times L_S \times L_S}$, and $A_T \in \mathbb{R}^{L_S \times L_T \times L_T}$ to $\tilde{A}_S \in \mathbb{R}^{L \times L}$ and $\tilde{A}_T \in \mathbb{R}^{L \times L}$, via zero-padding and matrix dimension merging. Thus, we can derive Eq. (5) using Eq. (7) and the associative property of matrix multiplication as follows

$$\begin{aligned} Attn_{mix} &= w_{3D}\mathcal{S}(A_{3D})V_{3D} + w_S\mathcal{S}(\tilde{A}_S)V_{3D} + w_T\mathcal{S}(\tilde{A}_T)V_{3D} \\ &= [w_{3D}\mathcal{S}(A_{3D}) + w_S\mathcal{S}(\tilde{A}_S) + w_T\mathcal{S}(\tilde{A}_T)]V_{3D} \end{aligned} \tag{8}$$

where $\mathcal{S}(.)$ represents the Softmax function. By utilizing matrix extraction in Eq. (6) and weighted matrix summation in Eq. (8), we save the computational cost associated with calculating attention matrices and multiplying them with the value tensor for both spatial and temporal attention. In this manner, we are able to condense the training-time multi-branch structure to the inference-time plain structure, as depicted in Fig. 3(A).

The pipeline of inference-time branch fusion method is illustrated in Algorithm 1. Notably, in the algorithm we avoid using the matrices $\tilde{A}_S$ and $\tilde{A}_T$ and instead sum the matrices A_S and A_T directly with the corresponding elements of A_{3D} to optimize memory usage.

Algorithm 1 Inference-time branch fusion.

Input: feature tensor $X \in \mathbb{R}^{L \times D}$, learned branch weights w_{3D}, w_S, w_T;
Output: mixed attention result $Attn_{mix}$;

1: Generate the query, key, and value tensors from X:
   $Q_{3D}, K_{3D}, V_{3D} \leftarrow X[W_q, W_k, W_v] \in \mathbb{R}^{L \times D}$;
2: Compute the 3D attention matrix:
   $A_{3D} \leftarrow Q_{3D}K_{3D}^T/\sqrt{D} \in \mathbb{R}^{L \times L}, L = L_T \times L_S$;
   $A'_{3D} \leftarrow Reshape(A_{3D}) \in \mathbb{R}^{L_T \times L_T \times L_S \times L_S}$;
3: Extract the spatial attention matrix:
   $A_S \leftarrow A'_{3D}[range(L_T), range(L_T), :, :]$;
   $A_S \in \mathbb{R}^{L_T \times L_S \times L_S}$;
4: Extract the temporal attention matrix:
   $A_T \leftarrow A'_{3D}[:, :, range(L_S), range(L_S)]$;
   $A_T \in \mathbb{R}^{L_S \times L_T \times L_T}$;
5: Apply the softmax function:
   $A_{3D} \leftarrow Reshape(A'_{3D}) \in \mathbb{R}^{L \times L}$;
   $A_{\{3D,S,T\}} \leftarrow Softmax(A_{\{3D,S,T\}})$;
   $A'_{3D} \leftarrow Reshape(A_{3D}) \in \mathbb{R}^{L_T \times L_T \times L_S \times L_S}$;
6: Compute the weighted sum of three attention matrices:
   $A'_{3D} \leftarrow w_{3D}A'_{3D}$;
   $A'_{3D}[range(L_T), range(L_T), :, :] \mathrel{+}= w_S A_S$;
   $A'_{3D}[:, :, range(L_S), range(L_S)] \mathrel{+}= w_T A_T$;
   $A_{3D} \leftarrow Reshape(A'_{3D}) \in \mathbb{R}^{L \times L}$;
7: Compute the mixed attention result:
   $Attn_{mix} \leftarrow A_{3D}V_{3D} \in \mathbb{R}^{L \times D}$;
8: **return** $Attn_{mix}$;

3.3. STA-3DA Transformer

The proposed STA-3DA module is applicable to various backbone networks, which can serve as a replacement for 2D attention in image Transformers, as well as for 3D attention in video Transformers. For the sake of clarity, we present the STA-3DA Transformer with the ViT-B backbone, as shown in Fig. 4.

The input video is symbolized by $V \in \mathbb{R}^{T \times H \times W \times 3}$, where H, W, and T represent the height, width, and length of the video respectively. We first divide the frames into $K \times K$ non-overlapping patches and get sequential tensor $\hat{V} \in \mathbb{R}^{L \times 3K^2}$, with $L = T(HW/K^2)$. Next, the tensor is subjected to linear mapping via an embedding layer $E \in \mathbb{R}^{3K^2 \times D}$ and added with the positional encoding E_{pos}

$$X^0 = \hat{V}E + E_{pos} \tag{9}$$

It is worth noting that we also prepend an extra learnable classification token to the sequence $\hat{V}E$, which is ommited for simplicity. Subsequently, We feed the embedding X^0 into N Transformer encoders to extract video features, with each encoder incorporating the proposed STA-3DA module and a multi-layer perceptron (MLP) block as follows

$$\hat{X}^n = STA\text{-}3DA(LN(X^{n-1})) + X^{n-1} \tag{10}$$

$$X^n = MLP(LN(\hat{X}^n)) + \hat{X}^n \tag{11}$$

where $LN(.)$ represents layer normalization (Ba et al., 2016). Finally, we utilize an MLP head to process the classification token from the last encoder, generating frame-level classification predictions. We then perform temporal average pooling to produce the video category.

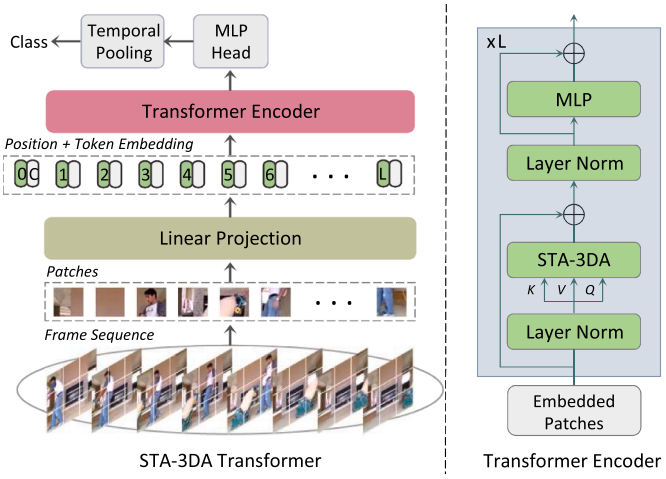

Fig. 4. Overview of the proposed STA-3DA Transformer. The input frame sequence is first partitioned into patches, which are then linearly projected and augmented with positional embeddings. The resulting sequence is processed by a series of Transformer encoders containing the proposed STA-3DA modules and MLP blocks.

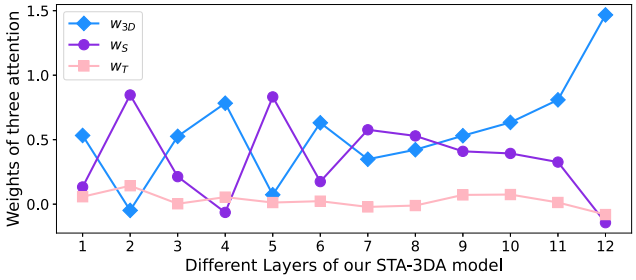

Fig. 5. Visualization of the branch weights in the STA-3DA model. This figure illustrates the learned weights of different attention branches across all layers.

4. Experiments

In this section, we first present the experimental setup, followed by conducting ablation experiments to analyze the impact of different configurations. Subsequently, we compare our approach with the state-of-the-art approaches on publicly video datasets and provide some visualization results.

4.1. Experimental setup

Datasets. Kinetics-400 (Carreira & Zisserman, 2017) comprises 400 action categories, each with a minimum of 400 videos. As a commonly used video dataset, it encompasses a diverse range of video categories, such as single-person actions (e.g., crying), person-object interactions (e.g., playing organ), and person-person interactions (e.g., shaking hands). Following Wang et al. (2021), we train the proposed model on the training set (around 240k videos) and employ the validation set (around 20k videos) for evaluation. Something-Something V2 (Goyal et al., 2017b) consists of 174 human-objection interaction classes. These classes are annotated based on textual descriptions using templates like Dropping [something], which allows the learned model to better capture motion-related features. In accordance with Bertasius et al. (2021), we train our network on roughly 170K training samples and evaluate its performance on about 25K validation samples.

UCF101 (Soomro et al., 2012) has 13,320 instances from 101 action classes, categorized into five types, such as playing musical instruments and sports. Following Zhang et al. (2022), we adopt the first split, with 9537 and 3783 clips in the training and validation sets. EGTEA Gaze+ (Li et al., 2018) is a first-person-vision dataset, which includes 10,321 videos and covers 106 action categories taking place in kitchen environment, such as opening fridge. In line with previous works (Zhang et al., 2021), we also report the performance of the first split, where 8299 and 2022 videos are respectively utilized for training and evaluation. Moments in Time (Monfort et al., 2019) is a human-annotated collection of 800,000 3-second videos for understanding dynamic events. Each video is labeled with one of 339 verb classes, spanning humans, animals, objects, and natural phenomena. The dataset emphasizes the understanding of spatiotemporal dynamics. AVA (Gu et al., 2025) is a densely annotated spatio-temporal atomic action dataset, with 211k training and 57k validation video segments. In line with the experimental protocol in SlowFast (Feichtenhofer et al., 2019), we present the mean Average Precision (mAP) results over 60 classes.

Baselines. We select the following state-of-the-art methods as baselines to evaluate our approach. (1) CNN-based methods, including 3D CNNs, e.g., I3D (Carreira & Zisserman, 2017), P3D (Qiu et al., 2017), S3D (Xie et al., 2018), Non-local (Wang et al., 2018), SlowFast (Feichtenhofer et al., 2019), SAP (Wang et al., 2020b), SmallBigNet (Li et al., 2020a), X3D (Feichtenhofer, 2020), and CorrNet (Wang et al., 2020a), and 2D CNNs, e.g., TSM (Lin et al., 2019), TEA (Li et al., 2020b), TEINet (Liu et al., 2020), TANet (Liu et al., 2021), TDN (Wang et al., 2021), and GC-TDN (Hao et al., 2022). (2) Transformer-based methods, e.g., ViT (Dosovitskiy et al., 2020), Tokshift (Zhang et al., 2021), VTN (Neimark et al., 2021), VidTr (Li et al., 2021), ViViT (Arnab et al., 2021), TimeSformer (Bertasius et al., 2021), LAPS (Zhang et al., 2022), LookupViViT (Koner et al., 2024), and STTM (Feng et al., 2024).

Implementation Details. We instantiate the proposed STA-3DA Transformers with the ViT-B (Dosovitskiy et al., 2020) and Visformer-S (Chen et al., 2021) backbones. We initialize them with the weights pretrained on the ImageNet dataset (Deng et al., 2009). The branch weights for 3D, spatial, and temporal attention are initialized as 0.5, 0.5, and 0.05, respectively. To evaluate our method, we present the results for inputs with various resolutions, such as the standard size of $224 \times 224 \times 8$, as well as higher temporal and spatial resolutions. During training, the learning rate varies linearly with the batch size (Goyal et al., 2017a) and differs across datasets. When the batch size is configured as 2, the initial learning rates are correspondingly 0.04, 0.03, 0.04, and 0.06 for the Kinetics-400, Something-Something V2, UCF101, and EGTEA Gaze+ datasets. It is worth noting that following Arnab et al. (2021), we apply regularization techniques such as label smoothing and mixup on Something-Something V2, but not on other datasets. In all experiments, the total number of epochs is set to 18, and we decrease the learning rate by a factor of 10 at the 10th and 15th epochs. The training hyperparameter settings are detailed in Table A.1 of the appendix. For inference, we sample 5 clips from the input video and employ 3 spatial crops (left, center, and right), resulting in a testing configuration of 3×5.

4.2. Ablation study

We report the results of the ablation experiments on Kinetics-400, including the Top-1 accuracy under the 3×5 setting and the computational cost measured in FLOPs. Notably, the models are trained under the same hyperparameters in all comparative experiments.

Impacts of adding spatial and temporal attention branches. Table 1 presents the performance of combining 3D attention with spatial and temporal attention. We can observe that the incorporation of temporal and spatial attention leads to improved recognition accuracy (↑0.4 % - ↑0.8 %)across different backbones and input resolutions. These findings suggest that the proposed method demonstrates favorable adaptability to various networks and exhibits robustness to input size variations. Furthermore, it appears that the boosts in performance tend to be more significant for higher spatiotemporal resolution inputs. This can be attributed to our method's ability to fully exploit the spatiotemporal information inherent in high-resolution videos.

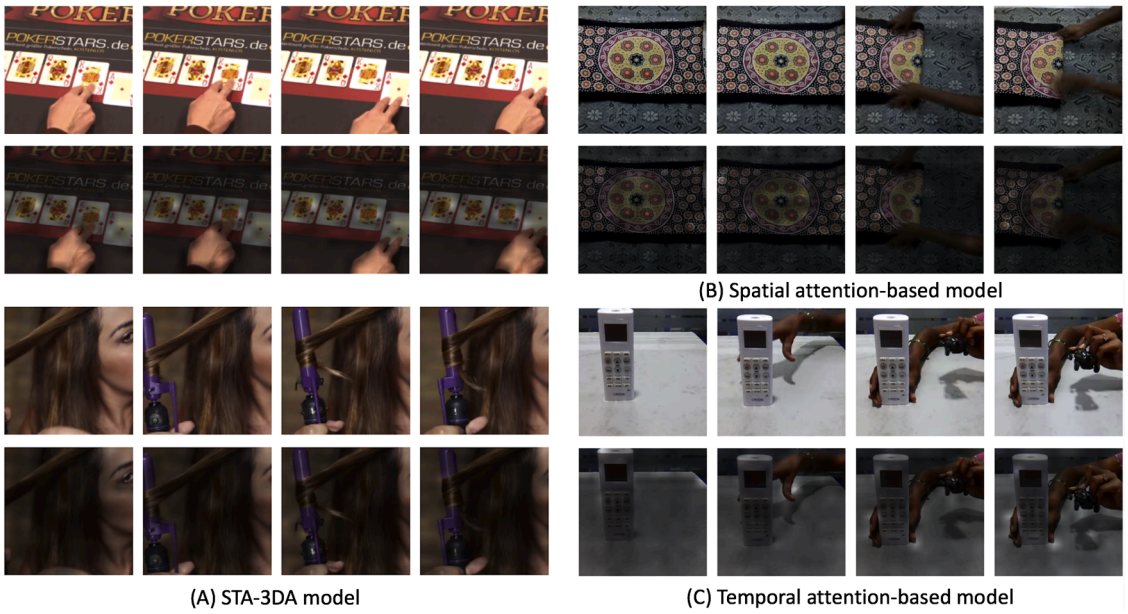

Fig. 6. Qualitative comparison of attention maps. The figure illustrates visualization results of the proposed STA-3DA, spatial-only, and temporal-only models on selected clips from Kinetics-400 and Something-Something V2 datasets, generated using Attention Rollout (Abnar & Zuidema, 2020). Input frames and their corresponding attention maps are displayed in odd and even rows, respectively.

Table 1
Impacts of incorporating spatial and temporal attention branches on top of 3D attention under different backbones and spatiotemporal resolutions.

Model	Backbone	Resolution	Top-1
3D	ViT-B	$224^2 \times 8$	78.7
3D + S	ViT-B	$224^2 \times 8$	78.9
3D + S + T	ViT-B	$224^2 \times 8$	**79.1** (↑0.4)
3D	ViT-B	$384^2 \times 8$	80.0
3D + S + T	ViT-B	$384^2 \times 8$	**80.6** (↑0.6)
3D	ViT-B	$224^2 \times 24$	79.9
3D + S + T	ViT-B	$224^2 \times 24$	**80.5** (↑0.6)
3D	Visf-S	$224^2 \times 8$	76.8
3D + S	Visf-S	$224^2 \times 8$	77.0
3D + S + T	Visf-S	$224^2 \times 8$	**77.2** (↑0.4)
3D	Visf-S	$360^2 \times 8$	78.5
3D + S + T	Visf-S	$360^2 \times 8$	**78.9** (↑0.4)
3D	Visf-S	$224^2 \times 24$	78.1
3D + S + T	Visf-S	$224^2 \times 24$	**78.9** (↑0.8)

Table 2
Impacts of the inference-time branch fusion operation on recognition accuracy and computational cost. "pre-F" and "post-F" separately stand for pre-fusion and post-fusion. The number of views (×15) are omitted in FLOPs.

Model	Backbone	FLOPs	Top-1
3D	ViT-B	181G	78.7
3D + S + T (pre-F)	ViT-B	187G	79.1
3D + S + T (post-F)	ViT-B	**181G**	**79.1**
3D	Visf-S	47G	76.8
3D + S + T (pre-F)	Visf-S	48G	77.2
3D + S + T (post-F)	Visf-S	**47G**	**77.2**

Table 3
Comparison of fixed and learnable weights for various attention branches.

Model	Backbone	Weights	Top-1
3D	ViT-B	-	78.7
3D + S	ViT-B	Fixed	78.7
3D + S	ViT-B	Learnable	**78.9**
3D + S + T	ViT-B	Fixed	78.6
3D + S + T	ViT-B	Learnable	**79.1**

Table 4
Comparison of various weight initialization methods for 3D, spatial, and temporal attention branches.

Model	Backbone	Initialization	Top-1
3D	ViT-B	-	78.7
3D + S + T	ViT-B	Equal	79.0
3D + S + T	ViT-B	Proportional	79.0
3D + S + T	ViT-B	Ours	**79.1**

Impacts of the inference-time branch fusion algorithm. Table 2 exhibits the effects of the branch fusion method shown in Algorithm 1 on accuracy and computational cost. "pre-F" and "post-F" represent the conditions pre- and post- branch fusion, respectively. The results indicate that regardless of using the ViT-B or Visformer-S backbone networks, the recognition performance remain consistent before and after the fusion process. This validates that by applying the structural re-parameterization strategy, the training-time multi-branch structure and the inference-time plain structure are computationally equivalent. Furthermore, our post-fusion model and 3D attention-based model have the same number of FLOPs, which signifies that the proposed STA-3DA module introduces negligible additional computational burden during inference compared to 3D attention.

Comparison of fixed and learnable branch weights. Table 3 compares the performance using fixed and learnable branch weights. "Fixed weights" refer to the weights of all encoders being the same and remaining unchanged during the training process, while "Learnable weights" means the branch weights of each encoder are trainable. We observe that using fixed weights does not improve the recognition results, whereas the addition of spatial attention and temporal attention with learnable weights lead to successive increases in accuracy.

Table 5
Comparison with the state-of-the-art methods on the validation set of the Kinetics-400 dataset.

Method	Backbone	Pretrain	Resolution	GFLOPs	Epochs	Top-1	Top-5
Two-Stream I3D (Carreira & Zisserman, 2017)	Inception V1	ImageNet	$224^2 \times 500$	216×NA	–	75.7	92.0
S3D-G (Xie et al., 2018)	Inception V1	ImageNet	$224^2 \times 250$	71×NA	112	74.7	93.4
Non-Local (Wang et al., 2018)	ResNet101	ImageNet	$256^2 \times 128$	359×30	196	77.7	93.3
SlowFast$_{16\times8}$ (Feichtenhofer et al., 2019)	ResNet101	None	$256^2 \times 32$	234×30	196	79.8	93.9
TSM (Lin et al., 2019)	ResNet50	ImageNet	$256^2 \times 16$	65×10	100	74.7	–
SmallBigNet (Li et al., 2020a)	ResNet101	ImageNet	$224^2 \times 32$	418×12	110	77.4	93.3
X3D-XL (Feichtenhofer, 2020)	X2D	None	$356^2 \times 16$	48×30	256	79.1	93.9
CorrNet (Wang et al., 2020a)	ResNet101	None	$224^2 \times 32$	224×30	250	79.2	–
TEA (Li et al., 2020b)	ResNet50	ImageNet	$256^2 \times 16$	70×30	50	76.1	92.5
TEINet (Liu et al., 2020)	ResNet50	ImageNet	$256^2 \times 16$	66×30	100	76.2	92.5
TANet (Liu et al., 2021)	ResNet50	ImageNet	$256^2 \times 16$	86×12	100	76.9	92.9
TDN (Wang et al., 2021)	ResNet101	ImageNet	$256^2 \times 24$	198×30	100	79.4	94.4
GC-TDN (Hao et al., 2022)	ResNet50	ImageNet	$256^2 \times 24$	110×30	100	79.6	94.1
MDAF (Wang et al., 2024b)	ResNet50	ImageNet	$224^2 \times 8$	34×30	150	76.2	92.0
ViT (Video) (Dosovitskiy et al., 2020)	ViT-B	ImageNet	$224^2 \times 8$	135×30	18	76.0	92.5
TokShift (Zhang et al., 2021)	ViT-B	ImageNet	$224^2 \times 16$	270×30	18	78.2	93.8
TokShift (MR) (Zhang et al., 2021)	ViT-B	ImageNet	$256^2 \times 8$	176×30	18	77.7	93.6
VTN (Neimark et al., 2021)	ViT-B	ImageNet	$224^2 \times 250$	4218×1	25	78.6	93.7
TimeSformer (Bertasius et al., 2021)	ViT-B	ImageNet	$224^2 \times 8$	197×3	15	78.0	93.7
TimeSformer-HR (Bertasius et al., 2021)	ViT-B	ImageNet	$448^2 \times 16$	1703×3	15	79.7	94.4
LAPS (Zhang et al., 2022)	Visf-S	ImageNet	$224^2 \times 8$	40×15	18	76.0	92.6
LAPS (L) (Zhang et al., 2022)	Visf-S	ImageNet	$320^2 \times 16$	173×15	18	78.7	93.8
LookupViViT (Koner et al., 2024)	ViT-B	ImageNet	$224^2 \times 32$	376×12	–	78.3	–
STTM (Feng et al., 2024)	ViT-B	ImageNet	$224^2 \times 11$	2288×1	30	80.2	–
STA-3DA	Visf-S	ImageNet	$224^2 \times 8$	47×15	18	77.2	93.0
STA-3DA	Visf-S	ImageNet	$224^2 \times 24$	191×15	18	78.9	94.0
STA-3DA	Visf-S	ImageNet	$360^2 \times 8$	147×15	18	78.9	93.9
STA-3DA	ViT-B	ImageNet	$224^2 \times 8$	181×15	18	79.1	94.2
STA-3DA	ViT-B	ImageNet	$224^2 \times 24$	817×15	18	80.5	94.7
STA-3DA	ViT-B	ImageNet	$384^2 \times 8$	788×15	18	**80.6**	**94.8**

Initialization strategy informed by weight analysis. Fig. 5 illustrates the learned weights of three attention branches in all STA-3DA Transformer encoders with the ViT-B backbone. We notice that the branch weights vary across different encoders, indicating that these encoders prioritize distinct visual elements. The utilization of learnable weights enables the model to dynamically capture noteworthy information in videos, which could be a potential reason for the superior performance of "Learnable weights" compared to "Fixed weights" shown in Table 3. It also can be seen that the weights of 3D and spatial attention are significantly higher than that of temporal attention. To verify the generality of this observation, we also analyze models with different backbones and training datasets, namely, a ViT-B model on Something-Something V2 and Visformer-S models on both Kinetics-400 and Something-Something V2. Consistent patterns emerge: when both 3D attention and spatial attention are used, the branch weight of 3D attention is marginally greater than that of spatial attention. When 3D, spatial, and temporal attention are combined, the weights of the 3D and spatial attention branches remain similar and are substantially larger than that of the temporal attention branch. This can be explained by the fact that 3D and spatial attention involve a greater number of token associations compared to temporal attention, thereby capturing richer visual semantic information and playing a more crucial role in the model's performance. Therefore, we initialize the branch weights close to the average of the learned weights across all encoders to promote model convergence during training. Concretely, we set both starting branch weights to 1/2 for combining only 3D and spatial attention, while the fusion of three attention involves allocating initial weights as 1/2, 1/2, and 1/20 respectively.

Sensitivity analysis of initialization strategy. Table 4 presents a comparison of different branch weight initialization methods. Here, The Equal strategy assigns same values to all weights (i.e., 1/3), and the Proportional strategy initializes the weights for 3D, spatial, and temporal attention proportional to their respective token counts (i.e., 1 : 1/8 : 1/196). The results show that both schemes improve recognition accuracy, indicating that model performance is not sensitive to this initialization. Nonetheless, our method (i.e., 1/2 : 1/2 : 1/20) achieves superior performance, as it promotes a more favorable optimization landscape during training.

Table 6
Comparison with the state-of-the-art methods on the validation set of the Something-Something V2 dataset.

Method	Backbone	Pretrain	GFLOPs	Top-1
TSM (Lin et al., 2019)	ResNet50	ImageNet-1K	65×6	63.4
GST (Luo & Yuille, 2019)	ResNet50	ImageNet-1K	59	62.6
STM (Jiang et al., 2019)	ResNet50	ImageNet-1K	67×30	64.2
SmallBigNet (Li et al., 2020a)	ResNet50	ImageNet-1K	157	63.3
SRATM (Xu et al., 2024)	ResNet50	ImageNet-1K	–	64.8
VidTr-L (Li et al., 2021)	ViT-B	–	351×30	63.0
TimeSformer-HR (Bertasius et al., 2021)	ViT-B	ImageNet-21K	1703×3	62.5
ViViT-L/16x2 (Arnab et al., 2021)	ViT-L	Kinetics-400	–	65.4
LookupViViT (Koner et al., 2024)	ViT-B	Kinetics-400	376×12	59.6
STA-3DA	Visf-S	Kinetics-400	191×15	**66.6**

4.3. Comparisons with the state-of-the-arts

Table 5 reports a comparison of the proposed model with various methods on the Kinetics-400 dataset, involving the backbones, the pretrained datasets, the spatiotemporal resolution of the input, the computational cost, as well as the Top-1 and Top-5 accuracy. The proposed model demonstrates superior performance compared to all the CNN-based methods listed including 3D CNN-based approaches (e.g., Two-Stream I3D (Carreira & Zisserman, 2017)) and 2D CNN-based approaches (e.g., TDN (Wang et al., 2021)). Furthermore, in comparison with other methods that adopt the ViT-B backbone, such as TokShift (MR) (Zhang et al., 2021), LookupViViT (Koner et al., 2024), and STTM (Feng et al., 2024), our approach attains better results. The proposed

Table 7
Comparison results on split 1 of UCF101.

Method	Backbone	Pretrain	Resolution	Top-1
P3D (Qiu et al., 2017)	ResNet50	Sports-1M	224^2×16	84.2
Two-Stream I3D (Carreira & Zisserman, 2017)	Inception V1	Kinetics-mini	224^2×500	96.5
TSM (Lin et al., 2019)	ResNet50	Kinetics-400	256^2×8	95.9
ViT (Video) (Dosovitskiy et al., 2020)	ViT-B	ImageNet-21K	256^2×8	91.5
TokShift (HR) (Zhang et al., 2021)	ViT-B	Kinetics-400	384^2×8	96.1
TokShift-L (HR) (Zhang et al., 2021)	ViT-L	Kinetics-400	384^2×8	96.8
LAPS (H) (Zhang et al., 2022)	Visf-S	Kinetics-400	320^2×32	96.9
STA-3DA	ViT-B	Kinetics-400	224^2×24	**97.4**

Table 8
Comparison results on split 1 of EGTEA-GAZE+.

Method	Backbone	Pretrain	Resolution	Top-1
TSM (Lin et al., 2019)	ResNet50	Kinetics-400	224^2×8	63.5
SAP (Wang et al., 2020b)	ResNet50	Kinetics-400	256^2×64	64.1
GC-TSM (Hao et al., 2022)	ResNet50	Kinetics-400	224^2×8	66.5
ViT (Video) (Dosovitskiy et al., 2020)	ViT-B	ImageNet-21K	224^2×8	62.6
TokShift (Zhang et al., 2021)	ViT-B	Kinetics-400	224^2×8	64.8
TokShift (HR) (Zhang et al., 2021)	ViT-B	Kinetics-400	384^2×8	65.8
LAPS (H) (Zhang et al., 2022)	Visf-S	Kinetics-400	320^2×32	66.1
STA-3DA	ViT-B	Kinetics-400	224^2×24	**68.2**

Table 9
Comparison results on the Moments in Time dataset.

Model	Pretrain	Top-1
TSN (Wang et al., 2016)	ImageNet-1K	25.3
TRN (Zhou et al., 2018)	ImageNet-1K	28.3
I3D (Carreira & Zisserman, 2017)	ImageNet-1K	29.5
AssembelNet-50 (Ryoo et al., 2020)	Kinetics-400	33.9
AssembelNet-101 (Ryoo et al., 2020)	Kinetics-400	34.3
ViT-3D (Dosovitskiy et al., 2020)	Kinetics-400	37.3
STA-3DA	Kinetics-400	**37.6**

Table 10
Comparison results on the AVA dataset.

Model	Pretrain	mAP
SlowFast, 4×16, R50 (Feichtenhofer et al., 2019)	Kinetics-400	21.9
SlowFast, 8×8, R101 (Feichtenhofer et al., 2019)	Kinetics-400	23.8
MViTv1-B, 16×4 (Fan et al., 2021)	Kinetics-400	24.5
ViT-3D (Dosovitskiy et al., 2020)	Kinetics-400	26.1
STA-3DA	Kinetics-400	**26.4**

method also surpasses the LAPS (L) model with the same Visformer-S backbone.

As shown in Table 6, we evaluate the proposed STA-3DA model on the Something-Something V2 dataset. Our model produces the best performance among these CNN-based and Transformer-based approaches. Moreover, when utilizing the weaker Visformer-S framework, our method incurs superior result and lower computational overhead than many other approaches (e.g., TimeSformer-HR (Bertasius et al., 2021) and LookupViViT (Koner et al., 2024)). The performance of our method on videos within this dataset that places emphasis on motion information validates its capability in learning motion dynamics.

To further assess the generalization ability of the proposed STA-3DA model pretrained on Kinetics-400, we conduct fine-tuning on small-scale video datasets. As presented in Table 7, our method delivers better performance on the UCF101 dataset than the listed CNN-based approaches, such as P3D (Qiu et al., 2017) and TSM (Lin et al., 2019). The proposed method also exhibits superior accuracy compared to some networks (e.g., TokShift-L (HR) (Zhang et al., 2021)) utilizing stronger backbones (e.g., ViT-L). We can also observe the similar results on the EGTEA-GAZE+ dataset in Table 8 that the STA-3DA model outperforms SAP (Wang et al., 2020b) and GC-TSM (Hao et al., 2022) based on ResNet50, and surpasses some given Video Transformers including TokShift (HR) (Zhang et al., 2021) and LAPS (H) (Zhang et al., 2022) by an improvement of over 2%.

Experimental results on the Moments in Time dataset are summarized in Table 9. This dataset presents considerable challenges for visual modeling due to high intra-class diversity, temporal sensitivity of actions, and significant label noise. Despite these difficulties, our method achieves superior recognition accuracy, outperforming both convolutional neural network-based approaches and the 3D attention-based ViT-3D model (Dosovitskiy et al., 2020). These results demonstrate the robustness of our method in understanding complex video content. Furthermore, as shown in Table 10, our method also obtains leading action detection results on the AVA dataset with a center testing strategy, surpassing strong competitors such as SlowFast (Feichtenhofer et al., 2019), MViT-B (Fan et al., 2021), and ViT-3D (Dosovitskiy et al., 2020). This consistent performance across different tasks and datasets highlights the generalizability and effectiveness of our method for downstream video understanding applications.

4.4. Visualization

Visualizing learned attention maps. In Fig. 6, we visualize the proposed STA-3DA model, the spatial and temporal attention-based models on some clips from the Kinetics-400 and Something-Something V2 datasets using Attention Rollout (Abnar & Zuidema, 2020). The attention matrices of different layers are recursively multiplied to propagate attention from the output token to the input space. The findings demonstrate that our STA-3DA module effectively captures regions with rich textures and dynamic motion patterns. For instance, in the "Playing cards" example, our model attends to both the performer's hands (motion) and the playing cards (object). Moreover, the spatial attention-based model selectively attends to salient foreground objects (e.g., the sajjada being folded), whereas the temporal attention-based model prioritizes motion-relevant areas (e.g., the trajectory of hand movements). By incorporating both spatial and temporal attention branches on top of the 3D attention mechanism, our model gains a strengthened ability to emphasize critical visual information in videos, namely the data pertaining to key targets and motion regions.

Visualizing learned video features. In Fig. 7, we present the visualization of the feature embeddings on EGTEA Gaze+ with t-SNE (Van der Maaten & Hinton, 2008). The various colored dots in the figure represent videos belonging to different categories. After pretraining on Kinetics-400, we observe that the proposed model learns video representations that are more dispersed and easier to classify. This is consistent with the improvement in recognition accuracy of our method on this dataset brought by model pretraining technique.

Visualizing per-frame prediction scores. In Fig. 8, we illustrates a comparison of the predicted probabilities for each frame between our STA-3DA and 3D attention with the ViT-B backbone. The frames in each row are sequentially sampled from the same video in Kinetics-400. Our method exhibits superior performance than the ViT-3D model on these videos, showcasing its strong modeling capabilities.

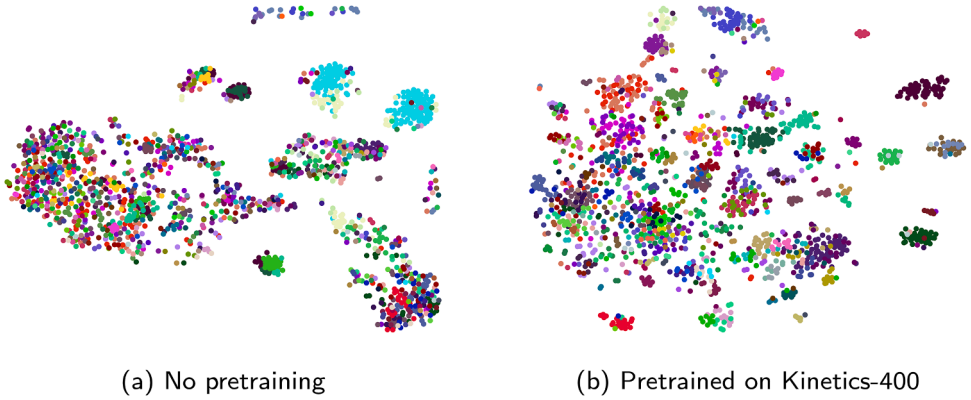

Fig. 7. t-SNE visualization of STA-3DA features on the EGTEA Gaze+ dataset, showing more dispersed and class-discriminative embeddings after Kinetics-400 pretraining.

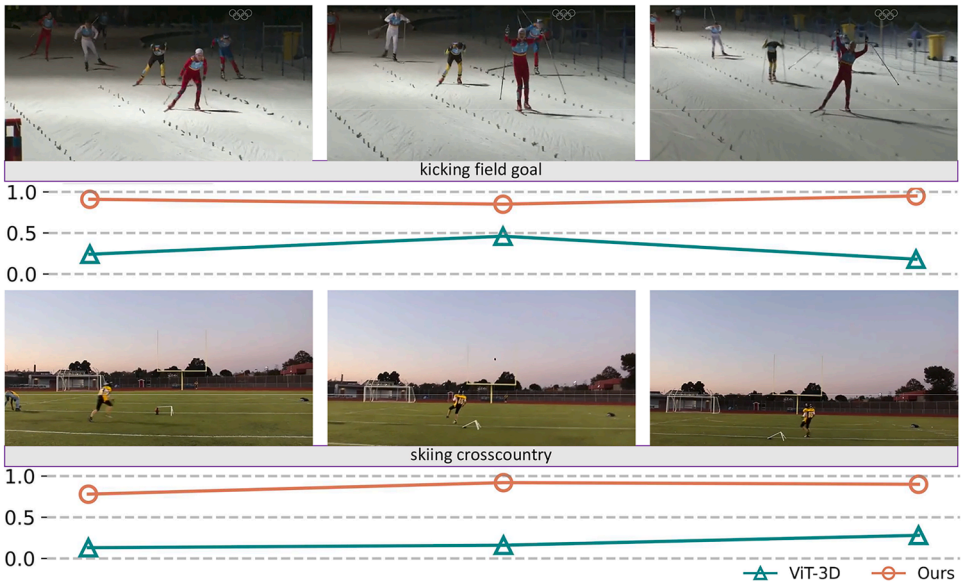

Fig. 8. Visualization of per-frame prediction scores on example videos from Kinetics-400, comparing the proposed STA-3DA model with ViT-3D.

5. Conclusion

This paper proposes a novel STA-3DA module, which integrates three complementary branches: 3D attention, spatial attention, and temporal attention. By incorporating spatiotemporal coherence prior, the design adaptively enhances token dependencies across both space and time, significantly boosting the model's representational capacity. Furthermore, we introduce a branch fusion algorithm based on structural reparameterization, enabling the seamless integration of spatial and temporal attention into the 3D attention pathway during inference. This results in negligible additional testing cost compared to standard 3D attention mechanism. Extensive experiments on large-scale public datasets, such as Kinetics-400 and Something-Something V2, demonstrate the effectiveness and efficiency of the proposed approach for video recognition and detection tasks. Additionally, the enhanced spatiotemporal representation learning of our method enables its application to tasks requiring fine-grained localization, such as spatiotemporal action detection. The architectural flexibility of STA-3DA also allows for extension to multimodal tasks (e.g., video captioning) through the integration of modality-specific attention mechanisms, which facilitates joint learning of cross-modal spatiotemporal alignments. To further enhance its capability and applicability, our future work will focus on generalizing the STA-3DA module to video understanding tasks beyond recognition and detection, along with incorporating more structural priors such as locality.

CRediT authorship contribution statement

Xiusheng Lu: Writing – original draft, Visualization, Validation, Software, Investigation, Data curation; **Lechao Cheng:** Writing – review & editing, Resources, Funding acquisition, Conceptualization; **Sicheng Zhao:** Writing – review & editing, Supervision, Methodology, Investigation, Formal analysis; **Ying Zheng:** Writing – review & editing, Investigation, Data curation, Conceptualization; **Yongheng Wang:** Writing – review & editing, Supervision, Resources, Funding acquisition; **Guiguang Ding:** Writing – review & editing, Supervision, Investigation, Funding acquisition, Formal analysis; **Mingli Song:** Writing – review & editing, Supervision, Resources, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Table A.1
Training hyperparamters on the Kinetics-400 and Something-Something V2 datasets, with "–" indicating that the regularisation method is not used.

	Kinetics-400	Sth-Sth v2
Optimisation		
Optimiser	SGD	SGD
Momentum	0.9	0.9
Batch size	2	2
Learning rate schedule	Step decay	Step decay
Decay step	1015	1015
Decay rate	0.1	0.1
Base learning rate	0.04	0.03
Epochs	18	18
Regularisation		
Random crop probability	1	1
Scale jitter probability	1	1
Random flip probability	0.5	–
Label smoothing λ	–	0.3
Mixup α	–	0.3

Acknowledgments

This work is funded by the Key R&D Program of Xinjiang, China under Project No. 2022B01006, the Exploratory Research Project of Zhejiang Lab under Project No. 2022PG0AN01, the National Natural Science Foundation of China under Project No. 62106235 and 62106236, and the National Key R&D Program of China under Project No. 2022YFE0137800.

Appendix A. Training hyperparameters

In this appendix, we provide the training hyperparameters on the two main datasets (i.e., Kinetics-400 and Something-Something V2), including optimization and regularization hyperparameters, as shown in Table A.1. Notably, we do not employ any regularization methods beyond data augmentation on Kinetics-400, and random flipping is not used on Something-Something V2 as it can potentially interfere with motion modeling.

References

Abnar, S., & Zuidema, W. (2020). Quantifying attention flow in transformers. arXiv:2005.00928.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE international conference on computer vision* (pp. 6836–6846).

Ba, J.L., Kiros, J.R., & Hinton, G.E. (2016). Layer normalization. arXiv:1607.06450.

Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? In *Proceedings of the international conference on machine learning* (pp. 813–824).

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6299–6308).

Chen, Z., Xie, L., Niu, J., Liu, X., Wei, L., & Tian, Q. (2021). Visformer: The vision-friendly transformer. In *Proceedings of the IEEE international conference on computer vision* (pp. 589–598).

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 248–255).

Ding, X., Guo, Y., Ding, G., & Han, J. (2019). Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In *Proceedings of the IEEE international conference on computer vision* (pp. 1911–1920).

Ding, X., Xia, C., Zhang, X., Chu, X., Han, J., & Ding, G. (2021a). Repmlp: Re-parameterizing convolutions into fully-connected layers for image recognition. arXiv:2105.01883.

Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., & Sun, J. (2021b). Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 13733–13742).

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

Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., & Feichtenhofer, C. (2021). Multiscale vision transformers. In *Proceedings of the IEEE international conference on computer vision* (pp. 6824–6835).

Feichtenhofer, C. (2020). X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 203–213).

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 6202–6211).

Feng, Z., Xu, J., Ma, L., & Zhang, S. (2024). Efficient video transformers via spatial-temporal token merging for action recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *20*(4), 1–21.

Gao, G., Liu, Z., Zhang, G., Li, J., & Qin, A.K. (2023). Danet: Semi-supervised differentiated auxiliaries guided network for video action recognition. *Neural Networks*, *158*, 121–131.

Gao, L., Liu, K., & Guan, L. (2025). A discriminative multi-modal adaptation neural network model for video action recognition. *Neural Networks*, *185*, 107114.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., & He, K. (2017a). Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv:1706.02677.

Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M. et al. (2017b). The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision* (pp. 5842–5850).

Gu, C., Sun, C., Ross, D. A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R. et al. (2025). Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Hao, Y., Zhang, H., Ngo, C.W., & He, X. (2022). Group contextualization for video recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 928–938).

Haria, A., Subramanian, A., Asokkumar, N., Poddar, S., & Nayak, J.S. (2017). Hand gesture recognition for human computer interaction. *Procedia Computer Science*, *115*, 367–374.

Jiang, B., Wang, M., Gan, W., Wu, W., & Yan, J. (2019). Stm: Spatiotemporal and motion encoding for action recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 2000–2009).

Klaser, A., Marszałek, M., & Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. In *Proceedings of the british machine vision conference* (pp. 1–10).

Koner, R., Jain, G., Jain, P., Tresp, V., & Paul, S. (2024). Lookupvit: Compressing visual information to a limited number of tokens. In *Proceedings of the european conference on computer vision* (pp. 322–337). Springer.

Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., Li, H., & Qiao, Y. (2023a). Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(10), 12581–12600.

Li, Q., Xie, X., Zhang, J., & Shi, G. (2023b). Few-shot human–object interaction video recognition with transformers. *Neural Networks*, *163*, 1–9.

Li, X., Wang, Y., Zhou, Z., & Qiao, Y. (2020a). Smallbignet: Integrating core and contextual views for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1092–1101).

Li, X., Zhang, Y., Liu, C., Shuai, B., Zhu, Y., Brattoli, B., Chen, H., Marsic, I., & Tighe, J. (2021). Vidtr: Video transformer without convolutions. In *Proceedings of the IEEE international conference on computer vision* (pp. 13557–13567).

Li, Y., Ji, B., Shi, X., Zhang, J., Kang, B., & Wang, L. (2020b). Tea: Temporal excitation and aggregation for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 909–918).

Li, Y., Liu, M., & Rehg, J.M. (2018). In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision* (pp. 619–635).

Lin, J., Gan, C., & Han, S. (2019). Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE international conference on computer vision* (pp. 7083–7093).

Liu, Z., Luo, D., Wang, Y., Wang, L., Tai, Y., Wang, C., Li, J., Huang, F., & Lu, T. (2020). Teinet: Towards an efficient architecture for video recognition. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 11669–11676).

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2022). Video swin transformer. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3202–3211).

Liu, Z., Wang, L., Wu, W., Qian, C., & Lu, T. (2021). Tam: Temporal adaptive module for video recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 13708–13718).

Luo, C., & Yuille, A.L. (2019). Grouped spatial-temporal aggregation for efficient action recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 5512–5521).

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11), 2579–2605.

Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S.A., Yan, T., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C. et al. (2019). Moments in time dataset: One million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 502–508.

Neimark, D., Bar, O., Zohar, M., & Asselmann, D. (2021). Video transformer network. In *Proceedings of the IEEE international conference on computer vision* (pp. 3163–3172).

Qiu, Z., Yao, T., & Mei, T. (2017). Learning spatio-temporal representation with pseudo-3d residual networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 5533–5541).

Ryoo, M.S., Piergiovanni, A.J., Tan, M., & Angelova, A. (2020). Assemblenet: Searching for multi-stream neural connectivity in video architectures. *International conference on learning representations*.

Scovanner, P., Ali, S., & Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the ACM international conference on multimedia* (pp. 357–360).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

Soomro, K., Zamir, A.R., & Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402.

Tan, Z., Li, X., Wu, Y., Chu, Q., Lu, L., Yu, N., & Ye, J. (2024). Boosting vanilla lightweight vision transformers via re-parameterization. In *International conference on learning representations*.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 4489–4497).

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6450–6459).

Voronin, V., Zhdanova, M., Semenishchev, E., Zelenskii, A., Cen, Y., & Agaian, S. (2021). Action recognition for the robotics and manufacturing automation using 3-d binary micro-block difference. *The International Journal of Advanced Manufacturing Technology*, *117*, 2319–2330.

Wang, A., Chen, H., Lin, Z., Han, J., & Ding, G. (2024a). Repvit: Revisiting mobile cnn from vit perspective. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 15909–15920).

Wang, B., Chang, F., Liu, C., Wang, W., & Ma, R. (2024b). An efficient motion visual learning method for video action recognition. *Expert Systems with Applications*, *255*, 124596.

Wang, H., Kläser, A., Schmid, C., & Liu, C.L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, *103*, 60–79.

Wang, H., Tran, D., Torresani, L., & Feiszli, M. (2020a). Video modeling with correlation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 352–361).

Wang, L., Tong, Z., Ji, B., & Wu, G. (2021). Tdn: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1895–1904).

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the European conference on computer vision* (pp. 20–36).

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7794–7803).

Wang, X., Wu, Y., Zhu, L., & Yang, Y. (2020b). Symbiotic attention with privileged information for egocentric action recognition. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 12249–12256).

Wu, F., Wang, Q., Bian, J., Ding, N., Lu, F., Cheng, J., Dou, D., & Xiong, H. (2022). A survey on video action recognition in sports: datasets, methods and applications. *IEEE Transactions on Multimedia*, *25*, 7943–7966.

Xie, S., Sun, C., Huang, J., Tu, Z., & Murphy, K. (2018). Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the european conference on computer vision* (pp. 305–321).

Xu, Y., Wang, Z., & Zhang, X. (2024). Leveraging spatial residual attention and temporal markov networks for video action understanding. *Neural Networks*, *169*, 378–387.

Zhang, H., Cheng, L., Hao, Y., & Ngo, C.w. (2022). Long-term leap attention, short-term periodic shift for video classification. In *Proceedings of the ACM international conference on multimedia* (pp. 5773–5782).

Zhang, H., Hao, Y., & Ngo, C.W. (2021). Token shift transformer for video classification. In *Proceedings of the ACM international conference on multimedia* (pp. 917–925).

Zhou, B., Andonian, A., Oliva, A., & Torralba, A. (2018). Temporal relational reasoning in videos. In *Proceedings of the european conference on computer vision* (pp. 803–818).

Zhou, W., Lin, K., Zheng, Z., Chen, D., Su, T., & Hu, H. (2025). Drtn: Dual relation transformer network with feature erasure and contrastive learning for multi-label image classification. *Neural Networks*, *187*, 107309.