IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 January 2024, accepted 18 March 2024, date of publication 20 March 2024, date of current version 4 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3379885

RESEARCH ARTICLE

Temporal Shift Module-Based Vision Transformer Network for Action Recognition

KUNPENG ZHANG, MENGYAN LYU, XINXIN GUO, LIYE ZHANG, (Member, IEEE), AND CONG LIU, (Member, IEEE)

College of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China

Corresponding author: Liye Zhang (zhangliye@sdut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62001272; and in part by the Natural Science Foundation of Shandong Province, China, under Grant ZR2023MF015.

ABSTRACT This paper introduces a novel action recognition model named ViT-Shift, which combines the Time Shift Module (TSM) with the Vision Transformer (ViT) architecture. Traditional video action recognition tasks face significant computational challenges, requiring substantial computing resources. However, our model successfully addresses this issue by incorporating the TSM, achieving outstanding performance while significantly reducing computational costs. Our approach is based on the latest Transformer self-attention mechanism, applied to video sequence processing instead of traditional convolutional methods. To preserve the core architecture of ViT and transfer its excellent performance in image recognition to video action recognition, we strategically introduce the TSM only before the multi-head attention layer of ViT. This design allows us to simulate temporal interactions using channel shifts, effectively reducing computational complexity. We carefully design the position and shift parameters of the TSM to maximize the model's performance. Experimental results demonstrate that ViT-Shift achieves remarkable results on two standard action recognition datasets. With ImageNet-21K pretraining, we achieve an accuracy of 77.55% on the Kinetics-400 dataset and 93.07% on the UCF-101 dataset.

INDEX TERMS Action recognition, self-attention, temporal shift module, vision transformer.

I. INTRODUCTION

With the rapid advancement of computer technology and the widespread adoption of the internet, video data has witnessed a notable trend of growth in both scale and complexity. The richness and diversity of video data have elevated it to a crucial data type in various industries, finding extensive applications in fields such as video surveillance, sports analysis, and intelligent driving. In these applications, video understanding technology plays a pivotal role, enabling people to recognize, analyze, and comprehend various behavioral actions from vast video datasets, as exemplified in Fig. 1's action recognition. To better exploit the potential of video data, the effective recognition of behavioral actions within videos becomes paramount. However, the complexity and scale of video data present certain challenges to the research and application of video understanding technology, including issues pertaining to video data storage, processing, computation, analysis, and comprehension.

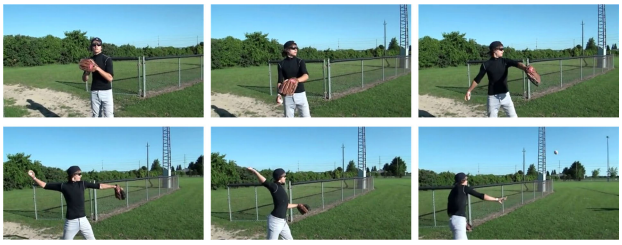


FIGURE 1. Catching or throwing a ball. The video encompasses a wealth of information, and through analysis of multiple video frames, it becomes evident that the individual within the frame is engaged in a pitching action.

In the past decade, the widespread application of convolutional neural networks (CNNs) has greatly enhanced the efficiency of problem-solving. Since the introduction of the

The associate editor coordinating the review of this manuscript and approving it for publication was Szidonia Lefkovits.

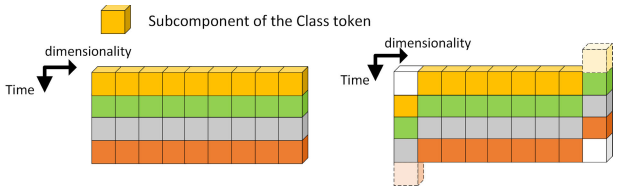


FIGURE 2. Shift module, with the left image depicting the original non-shifted state, and the right image showing the state after shifting. In our transformer, we only apply shifts to the cls token to achieve information exchange between video frames from different time intervals.

AlexNet network [1], computer vision tasks have experienced significant advancements, with CNNs playing a crucial role in image classification [1], [2], [3], [4], [5], object detection [6], [7], [8], semantic segmentation [9], [10], [11], and other domains, yielding remarkable results. Simultaneously, in the field of action recognition, the application of CNNs has also led to a gradual improvement in accuracy, contributing significantly to the advancement of this domain [12], [13], [14], [15], [16], [17], [18], [19]. Moreover, in recent years, action recognition work based on transformer models [21], [22], [23], [24] has been flourishing and has garnered widespread attention from both academia and industry.

Regarding video processing tasks, temporal modeling is of paramount importance. In convolutional neural networks (CNNs), to enhance the modeling capability of the temporal dimension, most methods adopt 3D convolutions [16], [25], [26], [27], enabling the network model to be applicable to video processing tasks. However, with the rise of transformer models in the field of natural language processing (NLP) [20], we have witnessed outstanding performance in handling text sequences. Similarly, in computer vision, video data can be perceived as longer sequences, which further extends the application of transformer models in video understanding. In fact, in recent years, the number of transformer-based video understanding approaches has been increasing, becoming one of the hot research topics in this field.

Our approach utilizes a transformer-based network for video processing. However, compared to handling individual images, processing video data requires greater computational resources. To mitigate the computational complexity, we draw inspiration from the seminal work of 2D CNN video processing, particularly TSM [29]. We incorporate shift modules to simulate temporal interactions, enabling more effective information exchange and reducing the computational burden on the temporal dimension. To provide a more intuitive representation, we perceive a video as a sequence of continuous T frames, where each frame can be denoted by $A \in R^{N\times C\times H\times W}$. Similarly, the video can be represented as $A \in R^{N\times C\times T\times H\times W}$). The schematic diagram of the Shift module is depicted in Fig. 2.

Our action recognition processing model is based on the Vision Transformer [30], which is originally designed for 2D image classification. In combination with the shift operation, we simulate the interaction process of temporal information, allowing the video frames entering the network to move bidirectionally in the channel dimension. This shift operation serves as a zero-computation module and can be easily embedded as a plug-in into 2D transformers, enabling various 2D transformer networks to be applicable for video understanding tasks.

Please note that our shift module does not move all tokens in the ViT, as doing so would incur significant data movement costs. Instead, we choose to move only the cls token, as based on the interpretation of ViT, the cls token can be considered as a token containing global information. Therefore, shifting the cls token can yield approximate effects while saving considerable resources and time. Furthermore, despite the shift module having zero computational cost, channel shifting comes with additional overhead. Thus, we only apply the shift operation before the multi-head attention, similar to the approach used in 2D CNN with shift modules [29]. This approach significantly alleviates the burden on hardware processing.

The contributions of our research work can be summarized as follows:

- Our proposed ViT-Shift action recognition model leverages the shift module to process videos, enabling the 2D Transformer network to handle video data more effectively. Moreover, this approach not only reduces computational complexity but also enhances the model's fitting accuracy.
- We carefully designed the utilization of the shift module, applying it only at critical positions, and experimental results have demonstrated the effectiveness of this approach. Compared to methods that apply shift operations at multiple positions, our strategy not only enhances the model's accuracy but also reduces the hardware burden.
- We solely utilize video RGB data as input and achieve impressive accuracy rates of 77.55% on the Kinetics-400 dataset and 93.07% on the UCF-101 dataset, with pretraining on ImageNet-21k. These results affirm the correctness of our approach, and we firmly believe that the shift module holds the potential for broad application and promotion in the field of computer vision, fostering further advancements in related research.

II. RELATED WORK

A. CONVOLUTIONAL NEURAL NETWORKS (CNN)

The Convolutional Neural Network (CNN) is a versatile and effective approach widely applied in computer vision, particularly for video recognition tasks [12], [14], [19], [32], [33]. Among them, the two-stream network stands as a classic architecture in deep learning [12]. Comprising two distinct network pathways, one dedicated to processing RGB inputs and the other to handling optical flow frames, these pathways interact at the network's final stage to yield video recognition results. This seminal work has inspired numerous subsequent studies [15], [31], [45]. Notably, Temporal Segment Networks (TSN) [15] introduced segment-based

sampling and aggregation modules, enabling the network to efficiently learn temporal information from distant time points in videos, significantly advancing the development of two-stream networks. Furthermore, the BS-2SCN network [45] introduced Bidirectional Gated Recurrent Units (BiGRU) and an attention mechanism based on neural science theory (SimAM), enhancing the spatial flow network's capability to extract features related to action appearance. This augmentation contributes to the improvement of the neural network's accuracy and stability. While two-stream networks can leverage results pre-trained on large-scale image datasets using 2D networks, handling optical flow data remains a cumbersome task.

With the advent of the C3D architecture [25], the research on 3D convolutional networks for video recognition entered a new era. C3D, a neural network featuring 3D convolutions, provided valuable insights for the subsequent development of 3D networks. The introduction of the I3D model [16] brought the concept of inflation to the forefront. The 3D network generated through inflating a 2D image classification network can address the issue of limited training data caused by the unavailability of the Imagenet dataset in C3D, thereby significantly enhancing video recognition accuracy. The SlowFast model [27] introduced a novel CNN architecture that combines two pathways—one specialized in processing temporal information and the other in spatial information. Their fusion within the network achieved state-of-the-art performance. X3D [19] is one of the best CNN models for video recognition, It is achieved through an iterative search process, utilizing a network search technique to explore various factors influencing video recognition. This approach yields optimal values for spatial-temporal resolution, feature channels, and network depth settings, leading to superior performance in video recognition tasks.

B. DECOMPOSED CONVOLUTIONAL NEURAL NETWORKS

To reduce the complexity of network models for video recognition, many researchers have focused on exploring methods to reduce network size while maintaining similar recognition performance. P3D [28] proposed a method that decomposes the convolution used for processing spatio-temporal information into two separate convolutions: one dedicated to processing spatial domain information independently and the other specialized in handling temporal domain information. This approach significantly reduces the complexity of the network and improves its computational efficiency. S3D [33] introduced a novel convolution strategy, where convolutions are performed separately along the temporal and spatial dimensions, replacing the 3D convolutions in the I3D network. This strategy greatly reduces the model's parameter count and enhances network performance. R (2+1) D [32] proposed a method of decomposing the 3D convolution kernel into independent spatial convolution kernels and temporal convolution kernels. This approach improves the network's recognition accuracy while reducing computational overhead. Compared to I3D, this method exhibits faster convergence and is easier to train. Additionally, Zhang et al. [49] introduced a residual network for optimization, decomposing the network structure of three-dimensional convolutional kernels into two types of kernels: spatial flow and temporal flow. They performed data flow fusion based on the two-stream network, effectively improving training rates.

C. NEXT-GENERATION VIDEO UNDERSTANDING MODELS

The Transformer was originally proposed in the field of natural language processing and was used for sequence processing tasks, showing promising performance in various tasks such as machine translation. Its self-attention mechanism enables effective handling of long-range dependencies in these tasks. In recent years, the Transformer has found widespread applications in natural language processing, computer vision, recommendation systems, and other domains. With the emergence of Vision Transformer (ViT) in the visual domain [30], the Transformer has gradually been applied to computer vision and demonstrated higher accuracy than traditional CNNs. Consequently, an increasing number of works have been built upon the Transformer architecture in computer vision, such as Cait [34], DeiT [35], and Swin Transformer [36], among others. These works have significantly improved the accuracy of image classification. Similarly, in the field of video classification, the introduction of VTN [24], VIVIT [21], TimeSformer [22], Tokshift [23], and LAPS [39] among others, has further facilitated the application of Transformers in video recognition tasks. Among them, TimeSformer [22] adopts a spatiotemporal decomposed attention mechanism, which calculates self-attention separately for the spatial and temporal dimensions of the video. This approach effectively reduces training and inference costs, leading to outstanding performance in video recognition tasks. VTN [24] conducts feature extraction for each frame in a video, and through temporal processing components, efficiently captures spatiotemporal dependencies in video sequences. This enhances the model's efficiency and accuracy, showcasing robust capabilities in video recognition tasks.

III. THE NETWORK MODEL

Our model draws inspiration from the TSM network [29], which has demonstrated excellent performance in the 2D CNN domain, achieving high network accuracy at a relatively lower computational cost. The pivotal role played by the shift module in the TSM network has motivated us to incorporate the shift module into the ViT network. The overall architecture of our network model is illustrated in Fig. 3. Next, in Section III-A, we will delve into the shortcomings of current action recognition models and introduce the advantages of our proposed new model. In Section III-B, we will provide a detailed description of ViT, the first transformer-based network applied to computer vision. Section III-C will focus on describing the shift

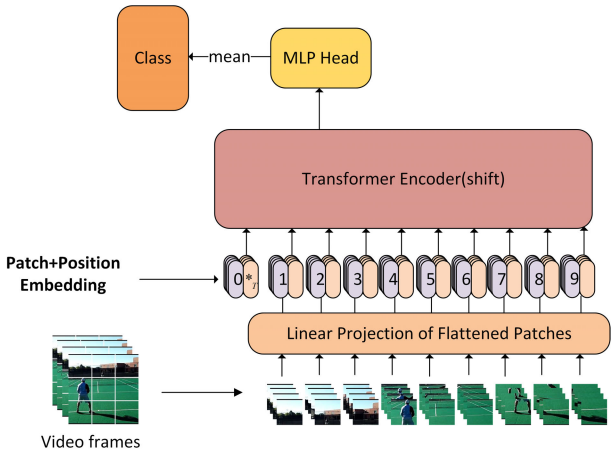


FIGURE 3. Overall architecture of our network model. In the design, we adhere to the principles of Vision Transformer while making efforts to retain its original structure and make necessary adjustments for video processing. After the tensor information is processed by Transformer (shift), it undergoes mean pooling before being utilized for prediction.

module, a component designed to reduce computational costs, and its application to video understanding tasks. Finally, in Section III-D, we will provide a comprehensive explanation of our new model, named ViT-Shift, which integrates ViT and the shift module and applies them to video action recognition tasks.

A. PROBLEM ANALYSIS

Currently, Transformer-based action recognition models have made significant progress in video understanding tasks. However, there are still some limitations. One crucial challenge is the modeling and interaction of temporal information. Traditional Transformer-based approaches typically use temporal attention mechanisms to handle temporal information in video sequences. However, this approach requires computing a large number of temporal attention weights, leading to high computational costs for the model.

To overcome these challenges and improve the performance of action recognition models, we propose an innovative Transformer-based action recognition model called ViT-Shift. Unlike previous methods, our model utilizes shift operations to simulate temporal interactions, avoiding the cumbersome computations of time-based attention. This shift operation efficiently captures temporal information in video sequences and significantly reduces computational complexity. In detail, we elaborate on the principles and applications of the shift operation in Section III-C. Additionally, the overall architecture of our model is depicted in Fig. 3, with further details available in Section III-D. By introducing shift operations, our model gains a better understanding and encoding of action patterns in videos, thereby enhancing accuracy.

Furthermore, compared to traditional convolution-based methods, our Transformer-based action recognition model exhibits significant advantages. Firstly, the Transformer model is capable of better modeling long-range dependencies, allowing the model to capture global contextual information in videos. Most importantly, our model demonstrates outstanding accuracy in action recognition tasks, fully validating its effectiveness and feasibility.

B. VISION TRANSFORMER (VIT) OVERVIEW

ViT [30] is regarded as one of the state-of-the-art models in the field of image classification, with its input being a single-frame image denoted as $X \in \mathbb{R}^{H\times W\times 3}$. ViT is an improved version based on the transformer architecture, achieved by removing the decoder part and retaining only the encoder part to focus on image recognition tasks. ViT comprises three core modules: the Embedding module, the Transformer Encoder module, and the MLP Head module.

1) EMBEDDING LAYER STRUCTURE

The primary function of the Embedding layer in ViT image processing is to divide the input image into N non-overlapping blocks of equal size, denoted as $N = \frac{H\times W}{Patch^2}$. Subsequently, each block $x_{img} = \mathbb{R}^{h\times w}$ is transformed into a one-dimensional vector Ex_i through linear projection. Before entering the Transformer encoder, position encoding is applied to the obtained multiple tokens, and they are concatenated with the class token representing global information. After these operations, the resulting tensor fed into the encoder is shown below:

$$z = [z_{cls}, Ex_0^1, Ex_0^2, \ldots, Ex_0^N] + P \tag{1}$$

In the above description, z_{cls} represents the token that focuses on global information, which plays a role in the final classification process. E denotes the encoding of each independent image block, mapping it to a one-dimensional vector of a specific dimension to maintain similarity with the original Transformer setup. x_0^N represents mutually independent segmented image blocks. Meanwhile, P represents the position encoding added to each token, ensuring that various image blocks transformed into one-dimensional vectors retain their spatial relationships in the original space.

2) TRANSFORMER ENCODER

After the Embedding module, N tokens are passed through L stacked Encoder Blocks. The Encoder Block primarily consists of Layer Norm, Multi-Head Attention, Dropout (or DropPath), and the final MLP Block. Layer Norm was originally proposed in the field of NLP, and with the introduction of Transformer, Layer Norm is also applied to normalize each token. Multi-Head Attention comprises multiple attention mechanisms, each separately calculating attention distributions. The calculation formula for each attention mechanism is as follows:

$$Attention(Q_i, K_i, V_i) = softmax(\frac{Q_i K_i^T}{\sqrt{d_k}})V_i \tag{2}$$

where Q_i represents a matrix composed of the query set, K_i represents a matrix consisting of a set of keys, V_i represents

a matrix composed of a set of values, and d_k represents the dimensionality of K_i.

After computing the attention for individual attention heads, a concatenation operation is performed. The specific computation formula is as follows:

$$\begin{aligned} Attention(Q, K, V) &= Concat(h_1, h_2, \ldots, h_m) \cdot W^o \\ h_i &= Attention(Q \cdot W_i^Q, K \cdot W_i^K, V \cdot W_i^v) \end{aligned} \tag{3}$$

where *Concat* denotes the concatenation operation applied to multiple attention heads. h represents each attention head, and m signifies the total number of attention heads. The computation process can be expressed as $Attention(Q \cdot W_i^Q, K \cdot W_i^K, V \cdot W_i^v)$, where W_i^Q, W_i^K, W_i^v are the corresponding learnable weight matrices, respectively.

3) MLP HEAD

Class token added at Embedding layer has global information. In the processing of the L Transformer Encoder, the cls token will participate in the computation of the attention mechanism as an independent token along with the token from the image block. In this interactive processing, the cls token will integrate the global image information and be used in the MLP Head to perform the classification task. The overall flow of ViT processing is shown below:

$$\begin{aligned} x^{li} &= MSA(DropPath(LN(z^l))) + z^l \\ z^{l+1} &= DropPath(MLP(LN(x^l))) + x^l \\ Class &= MLPHead(z^l) \end{aligned} \tag{4}$$

where *MSA* represents the Multi-Head Self-Attention module, capable of effectively processing image information. *DropPath*, as a regularization method, serves to enhance the model's expressive power. *LN* stands for Layer Norm, primarily employed for normalization operations to optimize the model's performance. *MLPHead* is a linear layer used for the final classification of image results. Specifically, the symbols are interpreted as follows: l denotes the number of stacked layers in the Encoder (the previous layer), L represents the total number of layers in the Encoder stack, and z signifies the tensor form of the processed image after transformation. *Class* denotes the final number of classification categories.

C. SHIFT MODULE

The Shift module is designed to simulate the interaction of temporal information and reduce computational complexity. Before processing with the Multi-Head Self-Attention (MSA) mechanism, the extracted temporal information from different video frames is fused. Specifically, this module achieves computational efficiency by combining the video frames at time steps t-1 and t+1 with the current frame at time t, thus reducing the significant computational overhead required by time-based attention. In the subsequent discussions, we will use the symbol *shift* to represent the Shift module.

$$c_l = shift(c_l) \tag{5}$$

Here, c_l represents the token containing global information.

We adopt the shift module that employs the idea of partial channel shifting for tokens. This approach is inspired by previous research [29] that utilized partial in-place shifting operations in 2D CNNs to achieve improved accuracy. Therefore, we interact the partial shift of the current time step with the preceding and succeeding frames while preserving the original semantic information of the unselected parts. However, the interaction process may lead to some loss of semantic information in certain parts. To address this, we use zero-padding operations to maintain the dimensionality of tokens unchanged.

We observed that performing shifting operations on all tokens at each time step t consumes a substantial amount of memory. Even though the shift operation itself is zero-computation, the significant delay overhead arises due to the extensive data movement, especially in large-scale data. To address this issue, we adopted the cls token used in ViT for aggregating global information. The cls token is set by the model during the initial training phase with random initialization. As training progresses, the cls token gradually integrates global information from the entire image through a multi-head self-attention mechanism, interacting with other tokens.

Specifically, in the self-attention mechanism, each token at a given position calculates its relevance to other tokens in the sequence, with these relevances represented as weights. Since the cls token is positioned at the start of the sequence and does not correspond to a specific local region, it can interact with each token that retains information about a specific local area of the image. This allows the cls token to serve as a central hub for global information, comprehensively considering various parts of the image in the self-attention mechanism. Through the processing of the multi-head self-attention mechanism, the cls token eventually fuses feature information from the entire image and encodes global semantic information into its representation.

This integration of global information makes the cls token a focal point for the model to holistically consider key elements of the entire image context during the inference stage, providing more accurate predictions for classification tasks. Since the cls token has already incorporated global information, simulating interaction effects between different time frames can be effectively achieved by shifting only the cls token.

So, we represent the cls token as c_l, which encapsulates the global information of the current frame. Our shift operation involves interacting the tokens at the current time step with the tokens from the previous and next time steps. Therefore, we split c_l into three parts along the channel dimension: c_{left}, c_{mid}, and c_{right}. Next, we perform shift interactions between the front and back parts of c_l with the adjacent two frames, while keeping the middle part unchanged. Here, we use the

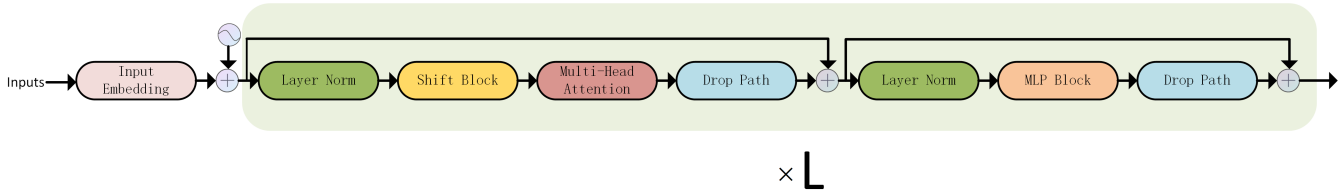


FIGURE 4. Model network architecture. The Shift Block represents the temporal shift module used for information interaction between different time frames. The encoder is stacked L times to process the data sequentially.

hyperparameter *div* as a switch to control the proportion of shifting. The detailed process is as follows:

$$\begin{aligned} c_l^t &= concat(c_{left}^t, c_{mid}^t, c_{right}^t) \\ c_{left}^t &= c_{left}^{t-1} \\ c_{mid}^t &= c_{mid}^t \\ c_{right}^t &= c_{right}^{t+1} \\ div &= \frac{c_{left}}{D}(or\frac{c_{right}}{D}) \end{aligned} \tag{6}$$

where t represents the current time step, $t-1$ represents the previous time step, $t+1$ represents the next time step, and D represents the number of channels in c_l.

In the previous section, we mentioned the properties of the shift operation, which is a zero-computation component that simulates the interaction of information and can save a significant amount of computation. However, too many shifts can lead to increased memory consumption and latency overhead, so caution is needed when using the shift module. The multi-head attention mechanism in ViT is similar to the convolution operation in neural networks, but it is more powerful. Therefore, we place the shift module only after the Layer Norm before the multi-head attention. Through experimental verification, we obtain satisfactory accuracy results, and detailed experimental results will be shown in Section IV. In addition, the setting of the shift ratio parameter is also our concern, and the appropriate shift ratio is crucial to the model performance. We will make comparisons in the ablation experiments to find more optimal shift ratio parameter settings.

D. VIT-SHIFT

The ViT structure is relatively straightforward, making it easy to train and adjust. Simultaneously, ViT possesses the capability to capture long-range dependencies in image sequences, exhibiting excellent generalization performance when handling large-scale data. This makes it a powerful tool for addressing complex visual tasks. The adaptability of ViT is particularly advantageous when applied to datasets for video action recognition.

In our work, we followed the design of the original ViT model and adapted it for the video understanding task. The main modifications were made to the input part of the model, and we added our shift component before the multi-head attention operation. The structure of our ViT-Shift network model is depicted in Fig. 4.

In the original ViT, the input consists of single-frame images, denoted as $X_{img} \in \mathbb{R}^{H\times W\times 3}$. However, videos are composed of multiple continuous frames of images. To better utilize the ViT architecture, we transform the video input $X_{video} \in \mathbb{R}^{T\times H\times W\times 3}$ into $X_{video} \in \mathbb{R}^{T\times N\times D}$ to adapt it to our variant of the ViT model, where H represents the height of the image, W represents the width of the image, T represents the input video frame length, $N = \frac{H\times W}{Patch^2}$, and D represents the channel dimension of each token.

In our model, all processed tokens are input into the encoder. Before entering the encoder, we add position encodings to each token to retain the original spatial position information of the video frames. Additionally, we introduce a global information cls token at this stage. Finally, the transformed input, adapted to the transformer format, is fed into the encoder for further processing.

In the encoder part, we perform repeated stacking of L layers on the encoder, and the dimension information of all tokens in the encoder is kept constant during the processing. By controlling the size of the value of L, our model can be manually adjusted to fit different hardware devices according to the actual application requirements. In each layer of the encoder, we first perform LN (Layer Normalization) on the tokens for normalization, and then introduce a shift module, which we designed to simulate the interaction of temporal information, for fusing information from different times in the subsequent multi-head self-attention operation. In order to prevent overfitting and increase the robustness of the model, we also applied the DropPath mechanism for random depth. Subsequently, all the tokens are again normalized by the LN and then mapped through the MLP module, which consists of two linear layers and a GELU activation function.

After undergoing L rounds of encoding, our model performs classification output through the MLP Head module. Below is a detailed processing flow of our model in video recognition tasks:

$$\begin{aligned} c^l &= shift(c^l) \\ z^l &= Drop\_path(MSA(LN(z^l))) + z^l \\ z^l &= Drop\_path(MLP\_block(LN(z^l))) + z^l \\ y &= \frac{1}{T}\sum_{i=1}^{T} MLP\_Head(c^l) \end{aligned} \tag{7}$$

where z^l represents all tokens, we average the class tokens of each video frame to perform the final action classification.

IV. EXPERIMENTAL EVALUATION

We evaluated our model on two standard datasets: the UCF-101 dataset [42] and the Kinetics-400 dataset [37]. As a baseline, we employed the original Vision Transformer architecture pretrained on ImageNet-21K. Our model is adaptable to various ViT sizes and exhibits excellent scalability. In Section IV-A, we provided a brief introduction to the datasets used. Next, in Section IV-B, we presented a detailed description of our model implementation. In Section IV-C, we conducted ablation studies to determine the optimal network hyperparameters. In Section IV-D, we present visualizations of our model's outcomes to underscore its prowess in capturing pivotal action-related insights. Lastly, in Section IV-E, we compared our model with other state-of-the-art approaches to demonstrate its feasibility and effectiveness.

A. DATASET

Kinetics-400 dataset [37]. The Kinetics-400 dataset is a large-scale and high-quality dataset of YouTube video clips aimed at encompassing a wide range of human-centered actions. This dataset consists of 400 human action categories, each containing at least 400 video clips. Each video clip has a duration of approximately 10 seconds and is sourced from various YouTube videos. The Kinetics-400 dataset covers a diverse set of action categories, including interactions between humans and objects (e.g., playing instruments) and interactions between individuals (e.g., shaking hands), as well as other relevant categories. As one of the standard datasets in the field of action recognition, its scale and quality provide a crucial benchmark for research in this domain.

UCF-101 dataset [42]. The UCF-101 dataset is an action recognition dataset of realistic action videos derived from YouTube, which contains 13,320 video samples from 101 different action categories. The dataset exhibits the greatest diversity in actions, covering significant differences in camera motion, object appearance and pose, object scale, viewpoint, background clutter, and lighting conditions. This makes the dataset challenging. Since most of the available action recognition datasets are mostly less than realistic and were constructed through staged participants, the goal of the UCF-101 dataset is to facilitate further research into the field of action recognition by learning and exploring new realistic action categories. The action categories in this dataset can be categorized into the following five types: 1) human-object interactions; 2) those involving only limb movements; 3) human-human interactions; 4) playing a musical instrument; and 5) physical activities.

We conducted experiments using the split-1 of the UCF-101 dataset and reported its performance metrics. Split-1 consists of 9,537 training videos and 3,783 validation videos. The details of the dataset information are provided in Table 1.

TABLE 1. Datasets. The Kinetics-400 dataset is published by [37]. The UCF-101 dataset is published by [42] and is shown with information from split-1 used in the experiment.

Dataset	Total	training	validation	categories
Kinetics-400	266,296	246,535	19,761	400
UCF-101	13,320	9,537	3,783	101

B. IMPLEMENTATION

Our model is developed based on the ViT architecture, making it applicable to various scales of ViT network models with excellent scalability. In our study, we chose to work with datasets comprising short videos lasting no more than 10 seconds, a common practice in the field. Short videos offer advantages in action recognition tasks as they are more adept at capturing essential information pertaining to specific actions, mitigating the presence of irrelevant noise compared to longer videos.

1) TRAINING

In our experiments, we used the ViT base architecture pretrained on the ImageNet-21K dataset. The learning rate of our model was set to 0.0609. We performed 8-frame sampling with a sampling interval of 32. To enhance the model's robustness, we adopted the same sampling preprocessing settings as in [23], including random resizing of video frames, random brightness adjustments, and horizontal flipping. For the Kinetics-400 dataset, we conducted training for 18 epochs, and for the UCF-101 dataset, we trained the model for 25 epochs. During training on the Kinetics-400 dataset, we reduced the learning rate by a factor of 10 at the 11th and 16th epochs, while for the UCF-101 dataset, we reduced the learning rate by a factor of 10 at the 10th and 20th epochs. Additionally, after conducting ablation studies on our model's most critical component, the shift module, we set the shift ratio to 4. Our model was trained on a Linux operating system using two Tesla T4 GPUs with 16GB of memory each. The CPU powering the machine is an Intel Xeon Gold 6242R.

2) TESTING

In our testing process, we followed the same evaluation strategy as in [23] to ensure a fair comparison with previous results. During testing, we sampled 10 video frames from each test video and resized each frame to the size of the shortest side. Then, we performed cropping from different directions on each video frame to generate 3 sub-frames of size 224×224. Finally, we averaged the predictions of the 30 video frames to obtain the final results of our network model. Furthermore, to ensure a fair comparison with various state-of-the-art models, we explored different numbers of sampled video frames. While keeping the spatial cropping settings consistent, we conducted experiments by sampling 1 frame and 3 frames from each test video, respectively. The detailed experimental results are presented in Table 2.

C. ABLATION EXPERIMENTS

To validate the actual impact of the Shift module on the final action classification, we conducted gain studies by

TABLE 2. Comparison with state-of-the-art techniques on Kinetics-400 Val. 'NA' and '-' indicates data not available.

Model	Pretrain	Inference Res (H × W)	Frames/Clip T	GFLOPs × Views	Params	Accuracy-1 %
ResLNet(V+L+B) [44]	None	112×112	64	NA×640	-	62.40
I3D [16]	ImageNet	224×224	250	108×NA	12M	71.10
R(2+1)D [32]	None	112×112	16	NA×10	33.3M	72.00
TSM [29]	ImageNet	256×256	8	33×10	24.3M	74.10
S3D-G [33]	ImageNet	224×224	250	71.3×NA	11.5M	74.70
GSF [43]	ImageNet	224×224	16	53.9×30	22.2M	74.80
GC-TSM [40]	ImageNet	224×224	8	33.3×10	25.6M	75.40
X3D-M [19]	None	256×256	16	6.2×30	3.8M	76.00
LAPS [39]	ImageNet-10k	224×224	8	40.1×15	39.8M	76.04
AGPN [47]	ImageNet	224×224	8	NA×30	27.6M	76.70
SlowFast8x8 [27]	None	256×256	32	65.7×30	-	77.00
Tokshift [23]	ImageNet-21k	224×224	8	134.7×30	85.9M	77.28
TimeSformer [22]	ImageNet-21k	224×224	8	590.0×3	121.4M	78.00
VTN [24]	ImageNet-21k	224×224	250	4218.0×1	114.0M	78.60
TP-ViT [50]	ImageNet	224×224	8	NA×NA	-	73.34
ARGC [51]	-	NA×NA	16	NA×NA	-	77.20
ViT-Shift(our)	**ImageNet-21k**	**224×224**	**8**	**134.7×1**	**85.9M**	**76.52**
ViT-Shift(our)	**ImageNet-21k**	**224×224**	**8**	**134.7×3**	**85.9M**	**77.07**
ViT-Shift(our)	**ImageNet-21k**	**224×224**	**8**	**134.7×30**	**85.9M**	**77.55**

adding the Shift module. Additionally, to optimize the performance of the Shift module, we conducted research on the hyperparameters controlling the shift ratio, aiming to find the most suitable hyperparameter settings for our network.

1) SHIFT MODULE

The shift module is the component responsible for implementing the shift operation. It facilitates information interaction along the temporal dimension. Without the use of shift operations, multiple frames input into the network would be treated as independent, lacking effective interaction with other frames. This would result in the loss of the video's dynamics and temporal context, potentially leading to a decline in the performance of video understanding. The shift operation addresses this issue, as described in Equation 6, by facilitating the interaction of partial information from the current frame with partial information from preceding and subsequent frames, while retaining essential information from the current frame. This enables the preservation of necessary spatial information while better integrating temporal information from different moments.

We conducted comparative experiments, comparing the ViT model with the added Shift module and the ViT model without the Shift module [23]. The results showed that the Shift module had a positive impact on video understanding tasks. On both datasets, the network model with the added Shift module outperformed the ViT model without the Shift module. The specific comparative results are shown in Fig. 5.

According to the data in Fig. 5, we observed that on the UCF-101 dataset, the addition of the Shift module resulted in a performance improvement of 1.61 percentage points compared to the ViT model without the Shift module. On the Kinetics-400 dataset, our model achieved a performance

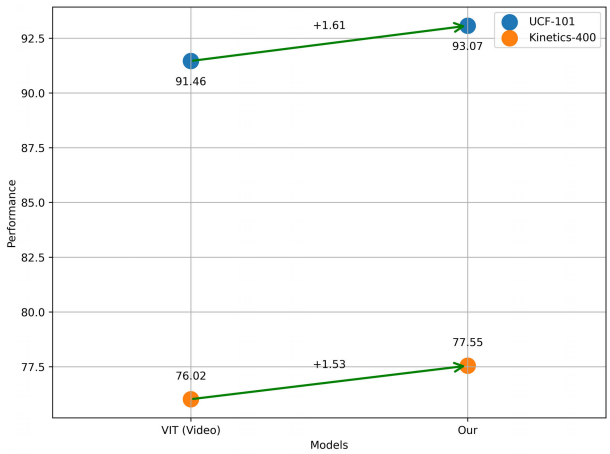


FIGURE 5. Comparison between using the shift module and not using the shift module.

improvement of 1.53 percentage points compared to the ViT model without the Shift module. These results are presented based on TOP-1 accuracy.

2) SHIFT RATIO SETTING

The shift ratio is a hyperparameter that controls the channel information interaction in the shift module. It determines the proportion of data in each token's feature vector that will undergo shifting. As mentioned in Equation 6, since c_l encompasses global temporal information, we exclusively shift c_l It is evenly divided into three parts along the channel dimension: left, center, and right. The left and right portions interact with the c_l representing global information from other moments, facilitating the results of temporal interaction. It is crucial to note that the shift ratio involves the amount of data exchange between the current moment c_l

and other moments c_l. Setting the shift ratio too high may result in the model losing too much spatial dimension information, making it challenging to distinguish different objects or scenes. Conversely, if the shift ratio is set too low, the information exchanged between different tokens will decrease, and the network model may fail to learn necessary temporal information, leading to an inability to capture changes in actions or objects. Therefore, finding a balance between retaining spatial and temporal information is essential to achieve the optimal video classification performance.

To comprehensively compare the impact of different shift ratios on model performance and identify the optimal parameter settings, we conducted ablation experiments on the UCF-101 dataset. The specific experimental results are detailed in Figure 6.

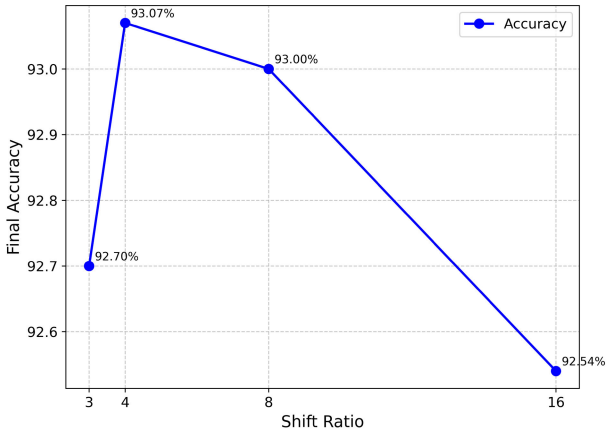


FIGURE 6. Comparison of shift ratios.

Experimental results have demonstrated that setting the shift ratio to 4 is the optimal choice for action recognition tasks, as it helps to enhance the performance of the model.

D. ATTENTION VISUALIZATION

The attentional effect of the network can be evaluated subjectively. In Fig. 7, we present the visualizations of our model's application of attention mechanism on video clips. Specifically, through visualizing attention, we gain a better understanding of the reasoning process of our model in video comprehension tasks and verify its focus on crucial actions or objects.

We utilized the visualization script provided in [30] to achieve the visualization functionality. Fig. 7 showcases video clips sampled from the Kinetics-400 dataset. In the first skiing clip, we can observe the model's focus on the skier, a behavior that aligns closely with human attention. In the second dog-walking clip, our model attends to both the person and the dog, effectively ignoring the surrounding environment. This further substantiates our network's ability to concentrate on the main subjects and perform accurate recognition.

FIGURE 7. Attention Visualization for Video Segments. The results indicate that our model pays particular attention to the relevant parts of the action subjects in the video during inference.

E. COMPARISON TO STATE-OF-THE-ART

1) COMPARISON ON THE KINETICS-400 DATASET

Our improved model not only enhances the performance of the ViT model in video recognition tasks but also demonstrates outstanding performance when compared to many state-of-the-art methods. Based on the results obtained from the ablation experiments, we compared the performance of our model with the state-of-the-art methods on the Kinetics-400 and UCF-101 datasets. These two datasets are standard datasets in the field of video recognition and have a high level of confidence. Specific comparison results for Kinetics-400 can be seen in Table 2.

Based on the comparison results presented, we observed that our model achieves comparable or even better accuracy than the state-of-the-art (SOTA) methods on the Kinetics-400 validation set. This finding validates the practical effectiveness of our approach, which combines the ViT network architecture with the shift module to achieve significant advancements in video recognition tasks. Moreover, the choice of introducing the shift module before the multi-head attention operation has also been validated. Furthermore, our model demonstrates excellent performance in terms of computational costs, as evident in the metrics compared in Table 2, including "Frames/Clip," "GFLOPs × Views," and "Params."

Specifically, compared to traditional general video recognition models (2D or 3D networks) such as AGPN [47], GC-TSM [40], and others, our model has achieved a significant improvement in accuracy. This indicates that the introduction of Transformer can accelerate the rapid development of video recognition. It's noteworthy that under

the same input video frame resolution (where each video frame is equivalent to a single image frame in a video), for example, I3D [16], S3D-G [33], our model requires the fewest training video frames and performs excellently. Additionally, under the same number of input video frames, as in TSM [29], Tokshift [23], our model, using the lowest image resolution (224×224), still produces satisfying results. Moreover, in comparison with the latest work, both TP-ViT [50] and our work utilize ViT networks for action recognition. However, the results indicate a clear superiority of our approach over TP-ViT when using the same number of frames and image resolution. When compared to the ARGC network, which also employs a transformer architecture for action recognition, our model achieves superior results with fewer frames.

By comparing the computational complexity and the number of inference views of our model with those of others, such as VTN [24], TimeSformer [22], and LAPS [39], we demonstrate the effectiveness of our model in reducing computational load. Specifically, compared to TimeSformer, our computational load is reduced by 77.2%, with only a 0.93% decrease in accuracy. Similarly, in comparison to VTN, our computational load is reduced by 99.7%, accompanied by an accuracy drop of only 2.08%. Furthermore, the shift operation involves zero computational cost and zero parameters. Our model utilizes the same number of parameters as the ViT [23] model, and when compared to models like VTN and TimeSformer, which also employ the Transformer architecture, it indicates that our model requires fewer parameters. In summary, through comparisons with state-of-the-art methods, the effectiveness of our work has been robustly demonstrated.

Furthermore, it is worth noting that the Tokshift [23] network, which shares similarities with our model structure, utilizes two shift modules, whereas our model incorporates only one shift module, yet still achieves favorable results. In TSM [29], it was highlighted that although shift modules are zero-computation and zero-parameter structures, the cost of data movement they introduce remains non-negligible. Therefore, relatively speaking, our network demonstrates higher efficiency in practical applications, as it requires fewer shift modules while maintaining competitive performance.

2) COMPARISON ON THE UCF-101 DATASET

UCF-101 is widely recognized as a classic standard dataset for action recognition tasks, and many research works evaluate the performance of network models on small-scale datasets. To verify the performance of our model on smaller-scale datasets, we compared it with other SOTA models on the UCF-101 dataset. The specific results are shown in Table 3.

According to the results in Table 3, the model employed in this study demonstrates outstanding performance on the UCF-101 dataset. In the field of deep learning, convolution-based methods like Two-stream [12], TSN [15], I3D [16], and P3D [28] have shown excellent performance in recent years. However, with the introduction of attention mechanisms, traditional convolutional methods have gradually lost their advantages. Additionally, the transformer-based shift model used in this study exhibits significant advantages compared to other networks employing attention mechanisms, surpassing the performance of the 3DRseNet50-CS [41] model. Compared to TokShift [23], our model demonstrates smoother running performance and higher accuracy. It is noteworthy that, relative to recent models such as VideoMAE [38], Kang et al. [48], BS-2SCN [45], HPER-Net [46], Conv Transformer [52], SVFormer [53], our ViT-shift model continues to perform well on small datasets like UCF-101. This emphasizes the robustness and adaptability of our model, especially in scenarios with limited data.

TABLE 3. Comparison with State-of-the-Art (SOTA) methods on UCF-101 dataset.

Model	Acc1(%)
P3D [28]	84.20
I3D [16]	84.50
Two-Stream [12]	88.00
Kang et al. [48]	89.66
3DRseNet50-CS [41]	89.90
BS-2SCN [45]	90.10
ResLNet(V+L+B) [44]	90.50
HPER-Net [46]	91.20
VideoMAE [38]	91.30
TokShift [23]	91.65
Zhang et al. [49]	91.70
TSN [15]	91.70
Conv Transformer [52]	86.10
SVFormer-B [53]	86.70
ViT-Shift(Our)	**93.07**

V. CONCLUSION

In this paper, we proposed a novel approach called ViT-Shift for action recognition, which is based on the Vision Transformer (ViT) model with the incorporation of a shift module. Compared to existing research, our model demonstrates significant improvements in both computational efficiency and accuracy for video action recognition. Notably, our model introduces a zero-parameter and zero-computation shift module solely before the multi-head attention of the encoder. As a result, Compared to other action recognition models, our model has less computational effort and higher accuracy. We have validated the effectiveness of our model on the Kinetics-400 and UCF-101 datasets, where it shows remarkable performance. Our future work will focus on further enhancing the accuracy of action recognition models while maintaining low computational overhead.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst. Conf. (NIPS)*, 2012, pp. 84–90.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[5] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, Long Beach, CA, USA, 2019, pp. 6105–6114.

[6] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1440–1448.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inform. Process. Syst.*, 2015, pp. 1–9.

[8] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[9] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Oct. 2015, pp. 234–241.

[11] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[12] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 1, no. 1, 2014, pp. 568–576.

[13] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.

[14] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-stream ConvNets," 2015, *arXiv:1507.02159*.

[15] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. van Gool, "Temporal segment networks for action recognition in videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2740–2755, Nov. 2019.

[16] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the Kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.

[17] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[18] M. Zolfaghari, K. Singh, and T. Brox, "ECO: Efficient convolutional network for online video understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 695–712.

[19] C. Feichtenhofer, "X3D: Expanding architectures for efficient video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 200–210.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, Dec. 2017, pp. 6000–6010.

[21] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "ViViT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6816–6826.

[22] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proc. ICML*, vol. 2, no. 3, 2021, p. 4.

[23] H. Zhang, Y. Hao, and C.-W. Ngo, "Token shift transformer for video classification," in *Proc. 29th ACM Int. Conf. Multimedia*, NY, NY, USA, Oct. 2021, pp. 917–925.

[24] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video transformer network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3156–3165.

[25] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[26] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.

[27] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6201–6210.

[28] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5534–5542.

[29] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 7082–7092.

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[31] Y. Zhu, Z. Lan, S. Newsam, and A. Hauptmann, "Hidden two-stream convolutional networks for action recognition," in *Proc. Asian Conf. Comput. Vis.*, vol. 11363. Cham, Switzerland: Springer, 2018, pp. 363–378.

[32] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.

[33] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 305–321.

[34] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 32–42.

[35] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[36] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[37] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.

[38] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," 2022, *arXiv:2203.12602*.

[39] H. Zhang, L. Cheng, Y. Hao, and C.-W. Ngo, "Long-term leap attention, short-term periodic shift for video classification," in *Proc. 30th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2022, pp. 5773–5782.

[40] Y. Hao, H. Zhang, C.-W. Ngo, and X. He, "Group contextualization for video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 918–928.

[41] Z. Yi, Z. Sun, J. Feng, and K. Jia, "3D residual networks with channel-spatial attention module for action recognition," in *Proc. Chin. Autom. Congr. (CAC)*, shanghai, China, Nov. 2020, pp. 5171–5174.

[42] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[43] S. Sudhakaran, S. Escalera, and O. Lanz, "Gate-shift-fuse for video action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10913–10928, Sep. 2023, doi: 10.1109/TPAMI.2023.3268134.

[44] T. Wang, J. Li, H.-N. Wu, C. Li, H. Snoussi, and Y. Wu, "ResLNet: Deep residual LSTM network with longer input for action recognition," *Frontiers Comput. Sci.*, vol. 16, no. 6, Dec. 2022, Art. no. 166334.

[45] Z. Wang, H. Lu, J. Jin, and K. Hu, "Human action recognition based on improved two-stream convolution network," *Appl. Sci.*, vol. 12, no. 12, p. 5784, Jun. 2022.

[46] S. Ren and M. Ding, "Heavy pose empowered RGB nets for video action recognition," in *Proc. 3rd Int. Conf. Consum. Electron. Comput. Eng. (ICCECE)*, Jan. 2023, pp. 382–387, doi: 10.1109/ICCECE58074.2023.10135328.

[47] Y. Chen, H. Ge, Y. Liu, X. Cai, and L. Sun, "AGPN: Action granularity pyramid network for video action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 3912–3923, Aug. 2023, doi: 10.1109/TCSVT.2023.3235522.

[48] K. Kang, S. Park, H. Park, D. Kang, and J. Paik, "Action recognition using multi-stream 2D CNN with deep learning-based temporal modality," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Las Vegas, NV, USA, Jan. 2023, pp. 1–3, doi: 10.1109/ICCE56470.2023.10043568.

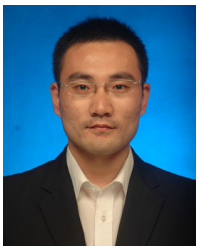
[49] J. Zhang and J. Zhao, "Dual-stream architecture and improved action recognition based on 3D convolutional neural network fusion," in *Proc. 4th Int. Conf. Comput. Eng. Appl. (ICCEA)*, Apr. 2023, pp. 181–185, doi: 10.1109/ICCEA58433.2023.10135414.

[50] Y. Jing and F. Wang, "TP-VIT: A two-pathway vision transformer for video action recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 2185–2189.
[51] Q. He and L. Shan, "Action recognition method based on graph neural network," in *Proc. 2nd Int. Conf. Mach. Learn., Cloud Comput. Intell. Mining (MLCCIM)*, Jiuzhaigou, China, Jul. 2023, pp. 144–148, doi: 10.1109/MLCCIM60412.2023.00027.
[52] N. H. Phong and B. Ribeiro, "Video action recognition collaborative learning with dynamics via PSO-ConvNet transformer," 2023, *arXiv:2302.09187*.
[53] Z. Xing, Q. Dai, H. Hu, J. Chen, Z. Wu, and Y.-G. Jiang, "SVFormer: Semi-supervised video transformer for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18816–18826.

KUNPENG ZHANG received the B.S. degree in computer science and technology from Shandong University of Technology, in 2022, where he is currently pursuing the master's degree in computer technology. His research interests include artificial intelligence, computer vision, and video comprehension.

MENGYAN LYU received the B.S. degree in network engineering from Zaozhuang University, in 2021. She is currently pursuing the master's degree with Shandong University of Technology. Her research interests include artificial intelligence, computer vision, image processing, and indoor positioning.

XINXIN GUO received the B.S. degree in computer science and technology from Shandong University of Technology, in 2021, where she is currently pursuing the master's degree in computer science and technology. Her research interests include artificial intelligence, computer vision, and simultaneous localization and mapping.



LIYE ZHANG (Member, IEEE) received the M.S., B.S., and Ph.D. degrees in information and communication engineering from Harbin Institute of Technology, Harbin, China, in 2009, 2011, and 2018, respectively. He is currently an Associate Professor with the School of Computer Science and Technology, Shandong University of Technology, and Harbin Institute of Technology. His research interests include indoor positioning and navigation, artificial intelligence, computer vision, visual information extraction, and robot application.

CONG LIU (Member, IEEE) received the bachelor's degree in engineering and the master's degree in computer software and theory from Shandong University of Science and Technology, in 2013 and 2015, respectively, and the Ph.D. degree in business data analytics from Eindhoven University of Technology, The Netherlands, in 2019. Currently, he holds the position of a Professor with the School of Computer Science and Technology, Shandong University of Technology. He has an extensive publication record in esteemed international conferences, including IEEE TRANSACTIONS, *Journal of Computer Science and Technology*, *Acta Automatica Sinica*, ICWS, ICSE, ICPC, WISE, EuroVIS, and KSEM. His research interests include business process management, process mining, data analytics, artificial intelligence, emergency management, and petri net theory and applications.