Intelligent Automation
& Soft Computing

Tech Science Press

DOI: 10.32604/iasc.2023.034855
Article

Fine-Grained Action Recognition Based on Temporal Pyramid Excitation Network

Xuan Zhou[1],* and Jianping Yi[2]

[1]School of Mechanical & Electrical Engineering, Xi'an Traffic Engineering Institute, Xi'an, 710300, China
[2]School of Electronics and Information, Xi'an Polytechnic University, Xi'an, 710048, China
*Corresponding Author: Xuan Zhou. Email: zhou_xuan668@163.com

Received: 29 July 2022; Accepted: 18 April 2023; Published: 23 June 2023

Abstract: Mining more discriminative temporal features to enrich temporal context representation is considered the key to fine-grained action recognition. Previous action recognition methods utilize a fixed spatiotemporal window to learn local video representation. However, these methods failed to capture complex motion patterns due to their limited receptive field. To solve the above problems, this paper proposes a lightweight Temporal Pyramid Excitation (TPE) module to capture the short, medium, and long-term temporal context. In this method, Temporal Pyramid (TP) module can effectively expand the temporal receptive field of the network by using the multi-temporal kernel decomposition without significantly increasing the computational cost. In addition, the Multi Excitation module can emphasize temporal importance to enhance the temporal feature representation learning. TPE can be integrated into ResNet50, and building a compact video learning framework-TPENet. Extensive validation experiments on several challenging benchmark (Something-Something V1, Something-Something V2, UCF-101, and HMDB51) datasets demonstrate that our method achieves a preferable balance between computation and accuracy.

Keywords: Fine-grained action recognition; temporal pyramid excitation module; temporal receptive; multi-excitation module

1 Introduction

The goal of action recognition is to analyze and represent the motion information of the human, to distinguish the behavior contained in the videos. Action recognition has received extensive attention from academia and industry due to its applications in intelligent monitoring, motion analysis, human-computer interaction, and so on. However, since video data often shows a strong complexity in the temporal dimension, building a general and efficient temporal module to learn compact spatiotemporal feature representations from videos remains a challenging topic.

Previous approaches [1–4] generally use the fixed temporal kernel to learn the local temporal context, then stack a large number of temporal convolutions to gradually increase the temporal receptive field of the network. However, these methods have the following problems: first, it is difficult for a single temporal kernel to learn complex motion representations when the amplitude of motion change is different. Secondly, the stacked method is computationally inefficient, and some discriminative temporal information may be greatly weakened in the transmission process of distant inter-frame information in the network. Transformer-based spatiotemporal models have made breakthroughs in vision tasks [5–9], which capture the long-term dependencies of all tokens on spatiotemporal positions via the self-attention mechanism. However, the redundant matrix multiplication introduces a lot of computational overhead, over parameterization often makes it computationally expensive and hard to train. In addition, some approaches [10,11] exploit complex multi-branch structures to extract multi-level temporal context in videos, but they still introduce unnecessary computation burden since both spatial convolution and temporal convolution are performed over all the feature channels. Recently, Duta et al. [12] proposed utilizing efficient pyramid convolution to enrich the spatial representation. Unfortunately, this method has rarely been discussed and implemented in video understanding.

To solve the above problems, we propose a Temporal Pyramid Excitation Network (TPENet) to effectively capture multi-level temporal context, which can achieve a good trade-off between computation and accuracy. The proposed TPENet consists of two cascaded components: Temporal Pyramid (TP) and Multi Excitation (ME) modules. To improve the efficiency of spatiotemporal representation learning, the proposed TP module splits the input channels in parallel and performs channel-wise temporal convolutions with different kernel sizes on each group to learn temporal semantics at different levels. The TPE module equipped with temporal pyramid convolution can expand the temporal receptive field and achieve short, medium, and long-term temporal context aggregation. On the other hand, the TP module ensures channel interaction sufficiency by channel shuffling, and without introducing extra computational cost. In addition, we proposed a Multi Excitation module to further improve the generalization performance of the network. Firstly, the CE module enables the model to perceive discriminative temporal information from refined feature excitations by integrating local temporal context into channel descriptors, thus further improving the temporal reasoning ability of the network. Then, the Channel-wise Temporal Excitation module learns the temporal importance weight on global spatial features. Finally, the Spatial-wise Temporal Excitation module learns the temporal relations of each pixel to represent the motion difference of different pixels. Our method can be flexibly integrated into 2D Convolutional Neural Networks (CNNs) in a plug-and-play fashion and trained in an end-to-end manner, thus significantly improving the baseline spatiotemporal representation in a plug-and-play fashion.

2 Related Work

2D CNNs-based. These methods learn spatiotemporal representations by embedding temporal modules into the backbone network. Some works [13,14] focus on adding temporal modules on the top layer of the network to complete the later temporal fusion, which captures the temporal relations between the frame-level feature maps. These methods tend to focus on coarse or long-term temporal structures, but cannot represents finer temporal relations in the local window. Other works are generally based on the sparse temporal sampling of Temporal Segment Networks (TSN) [15], and complete the temporal interaction through global embedding. For example, Lin et al. [16] proposed a temporal shift module, which moves some channels along the temporal axis to implicitly represent local temporal relations. To extract the motion information, Jiang et al. [17] proposed spatiotemporal and motion encoding, which extracted spatiotemporal and local motion features by using channel-wise spatiotemporal blocks and channel-wise motion blocks, respectively. Kwon et al. [18] proposed a motion squeeze block to establish the relation between adjacent frames and transform this relationship into motion features for subsequent prediction. Liu et al. [19] proposed a Temporal Enhancement-and-Interaction module to learn motion and channel-wise temporal representations. For efficient action recognition, some work proposed some efficient spatiotemporal encoding methods such as multi-view learning [20] and channel temporization operator [21], but it is hard to learn global temporal context. To capture the long-term temporal context, some works use multi-path convolutions [22] or global temporal attention [23] to learn complex video representations. However, the additional computational cost introduced by these methods is still not negligible.

3D CNNs-based. These methods perform spatiotemporal representation learning by extending 2D filters to the temporal dimension. Wang et al. proposed to decompose the spatiotemporal module into an appearance branch for spatial modeling and a relation branch for temporal modeling and to explicitly model the temporal relation by adopting the multiplicative interaction between pixels and filter responses across multiple frames. The vanilla 3D filter introduces mass computational cost, which leads to the overfitting of the network. Some work [24,25] reduces parameters by decomposing 3D kernel factors into the 2D kernel in space and 1D kernel in time. Other work [26,27] utilizes grouped convolution or depth-separable convolution to design compact 3D descriptors. Furthermore, Zhou et al. [28] proposed to integrate 3D convolution modules into 2D CNNs to reduce the complexity of spatiotemporal fusion. Zolfaghari et al. [29] proposed to use 2D CNNs at the bottom layer of the network to capture static spatial semantic information, and 3D CNNs at the top layer to process complex motion patterns between frames. However, 3D CNNs are difficult to capture long-range temporal dependencies due to their limited receptive fields. To tackle this challenge, we adopt a temporal pyramid excitation module to learn short, medium, and long-term temporal context. Besides, we propose our TPENet, achieving powerful performance for action recognition.

3 Temporal Pyramid Excitation Network

3.1 Temporal Pyramid Excitation Module

As discussed in the introduction, the previous methods are hard to learn complex motion representation with a fixed temporal kernel. To tackle such a problem, we propose an efficient Temporal Pyramid Excitation (TPE) module, by splitting the input features along the channel dimension and utilizing channel-wise temporal convolutions with different kernel sizes to expand the temporal receptive field.

Especially, given an input feature map, TPE sequentially performs a Temporal Pyramid (TP) module and a Multi Excitation (ME) module as illustrated in Fig. 1, to perform temporal modeling and enhance the TP module, respectively. The overall process can be summarized as:

$$\mathbf{Y} = \mathrm{Z}\left(\Psi\left(\mathbf{X}\right)\right) \otimes \mathbf{X} \tag{1}$$

where $\otimes$ denotes element-wise addition. Ψ and Z denote the TP module and ME module, respectively. TP module captures short, medium, and long-term temporal context information from the input feature maps through efficient temporal pyramid convolution. ME module aggregates temporal features of different levels by parallel multi-path excitation to emphasize the discriminative temporal information, thus improving the generalization ability of the model.

TP module. The vanilla spatiotemporal pyramid convolution significantly increases the computational costs of the video architecture since both spatial convolution and temporal convolution are performed over all the feature channels, thus we use channel splitting and reduce channel interactions to alleviate computing bottleneck. The proposed TP module is shown in Fig. 2.

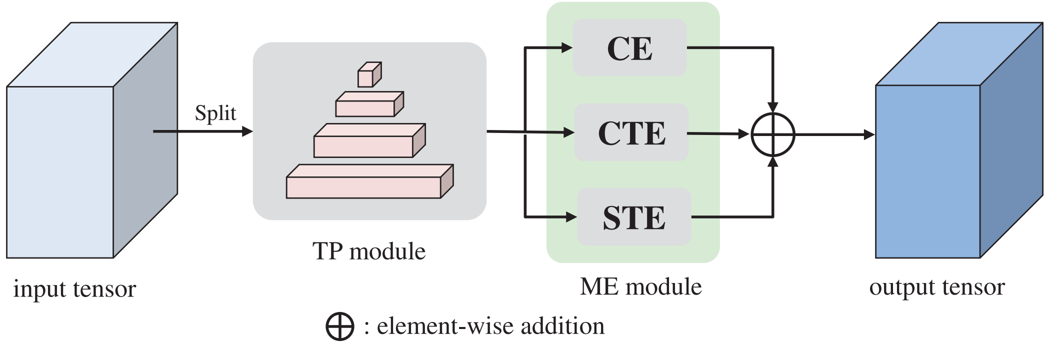

Figure 1: Temporal pyramid excitation module

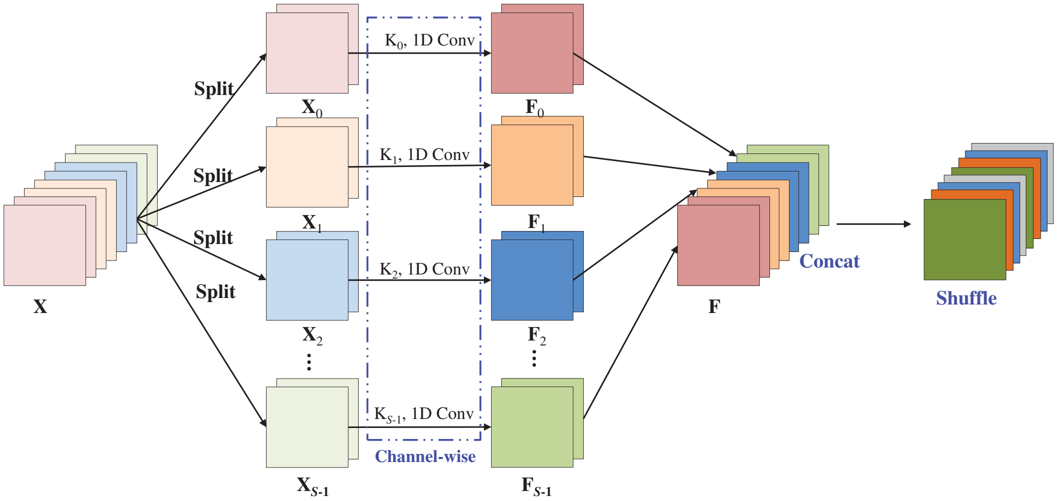

Figure 2: Temporal pyramid module

First, the input tensor $\mathbf{X}$ is split into S groups according to the channel dimension, namely $[\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_{S-1}]$, and each group has the same number of channels, where $\mathbf{X}_i \in \mathbb{R}^{T\times(C/S)\times H\times W}$.

Then, we employ temporal convolutions with different kernel sizes on the input feature map to perform efficient temporal modeling. Specifically, we first reshape input tensors $\mathbf{X}_i$ to $\mathbf{X}_i' \in \mathbb{R}^{HW\times C/S\times T}$, and use 1D channel-wise temporal convolution with different kernel sizes to generate temporal representation $\mathbf{F}_i \in \mathbb{R}^{HW\times C/4\times T}$ on multi-levels. This step can be explained as shown in Eq. (2).

$$\mathbf{F}_i = f_{K_i}(\mathbf{X}_i) \quad i = 0, 1, \cdots, S-1 \tag{2}$$

where, f_{K_i} denotes 1D channel-wise temporal convolution with kernel size K_i, $K_i = 2i - 1$.

Subsequently, we concatenate $\mathbf{F}_i$ according to the channel dimensions to aggregate the temporal contexts and reshape the output feature map $\mathbf{F}$ to $\mathbb{R}^{T\times C\times H\times W}$. This step can be shown in Eq. (3). In this way, it is beneficial for the network to learn multiple levels of temporal information from videos, thus improving the spatiotemporal representation ability for complex motion patterns.

$$\mathbf{F} = [\mathbf{F}_0, \mathbf{F}_1, \cdots, \mathbf{F}_{S-1}] \tag{3}$$

Finally, to overcome the impact of group convolution, we adopt the channel shuffle operation to ensure channel-interaction sufficiency.

In general, the proposed method differs from ordinary pyramid convolution in two aspects. (1) The original intention of our method is to improve the temporal receptive field rather than the spatial receptive field, (2) Our method improves the efficiency of temporal modeling by coordinating channel grouping and channel-wise convolution.

ME module. To effectively aggregate the multi-level temporal representation, we proposed the Multi Excitation (ME) module to explicitly model temporal importance weights. In general, ME contains three parallel paths, which can be further subdivided into Channel Excitation (CE), Channel-wise Temporal Excitation (CTE), and Spatial-wise Temporal Excitation (STE). The proposed three excitation modules are used to model the channel-temporal, temporal-only, and spatial-temporal relations, these joint modeling approaches can facilitate the network to extract complementary temporal features.

Channel Excitation (CE). To capture short-term temporal dynamics effectively, we propose a novel LTE module. LTE aims to learn local temporal relations by integrating temporal context information into channel descriptors so that the network can better focus on "what" and "when" for a given input image.

From Fig. 3, we first utilize global average pooling to aggregate the spatial information, thus obtaining a spatial global descriptor $\mathbf{F}_c \in \mathbb{R}^{T\times C\times 1\times 1}$, which can be represented as:

$$\mathbf{F}_c = \mathrm{GAP}(\mathbf{F}) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{F}(i,j) \tag{4}$$

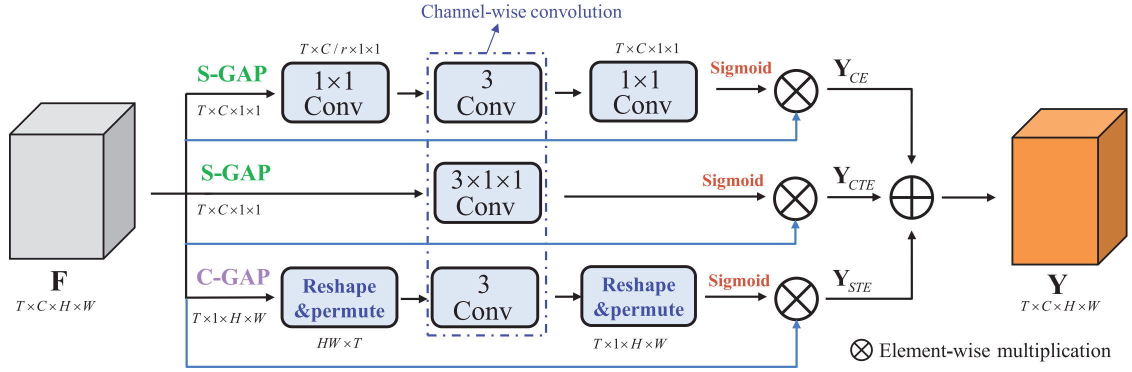

Figure 3: Multi-excitation module

Then, the global descriptor $\mathbf{F}_c$ is fed into a 1×1 convolution to squeeze the number of channels to C/γ ($\gamma = 16$). This step can be written as:

$$\mathbf{F}_c^* = \mathbf{W}_1 * \mathbf{F}^c \tag{5}$$

where $\mathbf{W}_1 \in \mathbb{R}^{C/r\times C}$, and $\mathbf{F}_c^* \in \mathbb{R}^{T\times C/\gamma\times 1\times 1}$. We then reshape the size of $\mathbf{F}_c^*$ to $\mathbb{R}^{C/\gamma\times T\times 1\times 1}$ to enable temporal modeling. A 1D channel-wise temporal convolution with kernel size 3 is utilized to integrate local temporal information into channel descriptors.

$$\mathbf{F}_t^* = \mathbf{K} * \mathbf{F}_c^* \tag{6}$$

where $\mathbf{K} \in \mathbb{R}^{3\times C}$ is parameterized by the kernel size and input channel, and $\mathbf{F}_t^* \in \mathbb{R}^{C/\gamma\times T\times 1\times 1}$. $\mathbf{F}_t^*$ is then reshaped to $\mathbf{F}_t \in \mathbb{R}^{T\times C/\gamma\times 1\times 1}$, which is then unsqueezed by using a 1 × 1 convolutional and fed to a Sigmoid activation. In this way, the channel attention mask $\mathbf{M}_c$ can be obtained, which can be formulated as:

$$\mathbf{M}_c = \delta\left(\mathbf{W}_2 * \mathbf{F}_t\right) \tag{7}$$

where $\mathbf{M}_c \in \mathbb{R}^{T\times C/\gamma\times 1\times 1}$, after the sigmoid function, each value of the tuple $\mathbf{M}_c$ is normalized to the interval [0,1] to measure the importance of each channel.

Finally, the output feature $\mathbf{Y}$ is obtained by performing channel-wise multiplication of the channel attention mask $\mathbf{M}_c$ and the input feature $\mathbf{F}$. The original features are recalibrated in the channel dimension by weighting the normalized weights to each channel. This process can be expressed as:

$$\mathbf{Y}_{CE} = \mathbf{F} \otimes \mathbf{M}_c \tag{8}$$

Channel-wise Temporal Excitation (CTE). To further capture efficient temporal information, we proposed a channel-wise temporal excitation module. In Fig. 3, we first use global spatial pooling to aggregate the spatial information, this step can refer to Eq. (4). Differing from CT-Net, we only use a channel-wise 3 × 1 × 1 convolution to obtain the temporal attention mask since channel reduction will lose the feature. Finally, we use the sigmoid function and the element-wise multiplication ⊗ broadcasts the temporal attention along the channel dimension, the whole process is shown in Eqs. (9) and (10).

$$\mathbf{M}_{CTE} = Sigmoid\left(Conv\left(S - GAP\left(\mathbf{F}\right)\right)\right) \tag{9}$$

$$\mathbf{Y}_{CTE} = \mathbf{F} \otimes \mathbf{M}_{CTE} \tag{10}$$

Spatial-Wise Temporal Excitation (STE). Previous temporal attention generally learns a local temporal attention mask on global spatial information, which complete disregard for spatiotemporal interactions can hurt performance since the motion directions of each foreground pixel are inconsistent. Using global spatial pooling to aggregate spatial information will make it difficult to learn the connection between spatial context and temporal context. Thus, we propose a spatial-wise temporal excitation as a complement to the CTE module that aims to learn the motion relations between pixels by explicitly modeling the element-wise temporal information. In Fig. 3, We first utilize 1D average pooling to average the channel responses to obtain the channel descriptors $\mathbf{F}' \in \mathbb{R}^{T\times 1\times H\times W}$. Then, we reshape and permute the shape of the channel descriptor from $\mathbb{R}^{T\times 1\times H\times W}$ to $\mathbb{R}^{HW\times T}$, and adopt 1D channel-wise temporal convolution to further learn the spatial-wise temporal importance weights $\mathbf{M}_{STE}$, this step can be formulated as:

$$\mathbf{M}_{STE} = \begin{Bmatrix} F'_{1,1} & F'_{1,2} & \cdots & F'_{1,T} \\ F'_{2,1} & F'_{2,2} & \cdots & F'_{2,T} \\ \vdots & \vdots & \vdots & \vdots \\ F'_{N,1} & F'_{N,2} & \cdots & F'_{N,3} \end{Bmatrix} * \begin{Bmatrix} w_{1,1} & w_{1,2} & w_{1,3} \\ w_{2,1} & w_{2,2} & w_{3,3} \\ \vdots & \vdots & \vdots \\ w_{N,1} & w_{N,2} & w_{N,3} \end{Bmatrix} \tag{11}$$

$$\mathbf{Y}_{STE} = \mathbf{F} \otimes \mathbf{M}_{STE} \tag{12}$$

where $*$ denote 1D channel-wise convolution. STE compresses local temporal information into channel descriptors to capture the temporal dependencies between pixels.

As shown in Fig. 3, we use simple element-wise addition to aggregate the output responses under different paths, which can be formulated as:

$$\mathbf{Y} = \mathbf{Y}_{CE} + \mathbf{Y}_{CTE} + \mathbf{Y}_{STE} \tag{13}$$

Discuss. TP module is designed similarly to the Timeception block. The differences are reflected in three parts: (1) TP module is inspired by the spatial pyramid convolution instead of the Inception block (Timeception). (2) For different groups of channels, the TP module utilizes different temporal convolutions instead of uniform temporal blocks to extract temporal context. (3) TP module reduces the computational cost by decomposing spatial/temporal and channel interactions, so it is significantly more efficient than the Timeception block.

3.2 Temporal Pyramid Excitation Network

The proposed Temporal Pyramid Excitation module can capture short, medium, and long-term temporal context information from videos, which is beneficial for the network to learn complex motion patterns.

In this section, we mainly explore the deployment of the TPE module in ResNet50. First, the TPE-block was established by embedding the TPE module into the basic bottleneck block of ResNet50. Then, we use TPE-block to replace all the basic residual units in the Conv_2 to Conv_5 of ResNet50, thus establishing an efficient spatiotemporal learning architecture, dubbed Temporal Pyramid Excitation Network (TPENet), which is illustrated in Fig. 4. Finally, all frame-level predictions are aggregated by temporal average pooling at the top of the network to obtain a video-level prediction score. Notably, the proposed TPE module only considers temporal information, the entire spatiotemporal modeling can be performed by utilizing the spatial modeling ability of 2D CNNs.

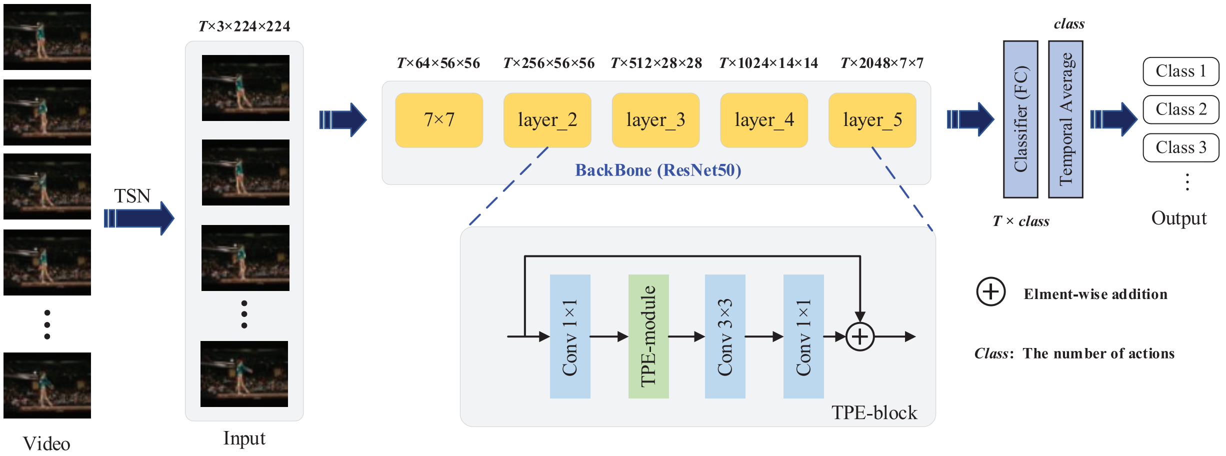

Figure 4: The overall architecture of the proposed TPENet

4 Experiment

4.1 Datasets and Evaluation Metrics

Since the goal of TPENet is to improve the inference ability of the network for long-term temporal structures and complex motion patterns, this paper evaluates the proposed method on Sth-Sth V1 [30], and other two small-scale datasets, UCF-101 [31] and HMDB-51 [32]. Among them, the Sth-Sth V1 dataset focuses on the fine-grained actions of human interaction with daily objects, including only the details of hand actions, and is used to emphasize the importance of temporal reasoning. This dataset contains 174 class actions with 108,499 video clips, including 86,017 training videos, 11,522 validation videos, and 10,960 test videos. Moreover, transfer learning experiments on UCF-101 and HMDB-51, which are much smaller than Kinetics and Sth-Sth, are carried out to verify the transfer capability of our solution.

We report top-1 and top-5 accuracy (%) on Sth-Sth V1 datasets. For UCF-101 and HMDB-51, we follow the original evaluation scheme using mean class accuracy. Also, we report the computational cost (in FLOPs) as well as the number of model parameters to depict model complexity.

4.2 Implementation Details

Training. We implement the proposed TPENet on Pytorch, and train the network via the mini-batch SGD, with an initial learning rate of 0.01, momentum is 0.9, and weight decay is 0.0005. We trained TPENet on four Nvidia RTX2080Ti, and sparsely sample 8 frames and 16 frames of 224 × 224 RGB images from the videos, the total batch size is set to 64 and 32 respectively. On the Sth-Sth V1 dataset, the network is trained for a total of 50 epochs. We use a cosine learning rate schedule, and the first 5 epochs are used for gradual warm-up to reduce the impact of mini-batch on model training. For UCF-101 and HMDB-51, we followed the common practice to fine-tune from Kinetics pre-trained weights and start training with a learning rate of 0.002 for 25 epochs. The learning rate decays 10 times every 10 epochs. Furthermore, we use a Dropout layer with a ratio of 0.5 after the global average pooling to alleviate over-fitting. Random scaling, multi-scale cropping, and flipping are used as data augmentation during training.

Test. If not specified, we use an efficient reasoning protocol to ensure reasoning speed. Firstly, 1 clip with T frames is sampled from a video. Then, each frame is resized to 256 × 256. Finally, the region of the center cropping is limited to 224 × 224 for action prediction.

4.3 Ablation Study

Deployment position. We first explore the impacts of different deployment locations by deploying TPE modules before the 1st convolution, before the 2nd convolution, before the 3rd convolution, and after the 3rd convolution in ResNet50, respectively. Fig. 5 shows different deployment positions, which are TPENet-a, TPENet-b, TPENet-c, and TPENet-d from left to right. The top-1 accuracy and FLOPs of different deployment positions are shown in Fig. 6.

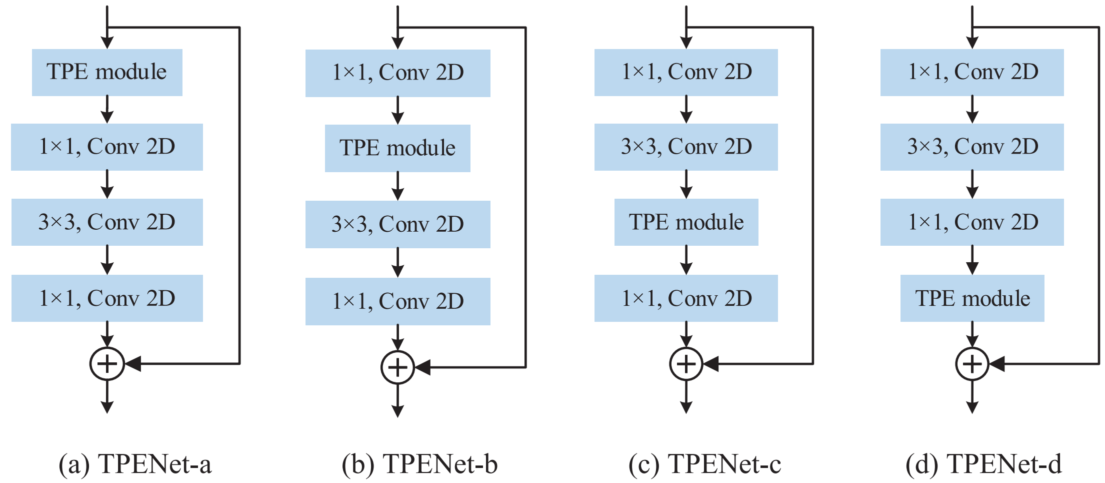

Figure 5: Deployment position of the TPE module in ResNet50

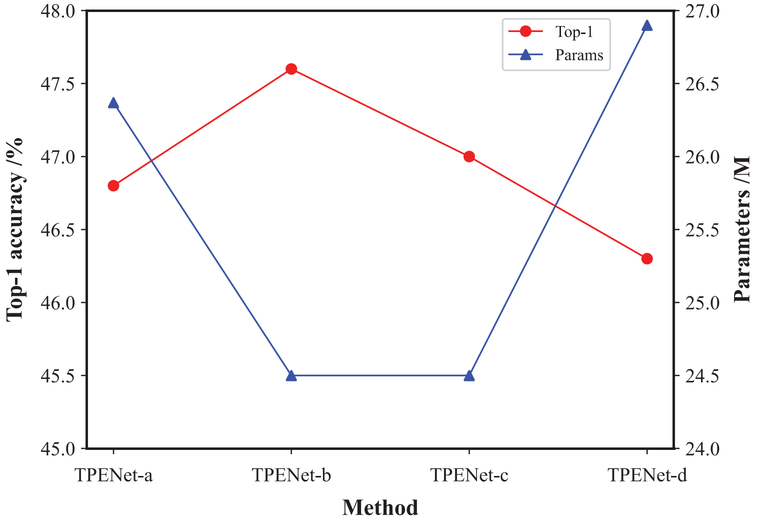

Figure 6: Accuracy and model complexity for different locations

From Fig. 6, the results show that top-1 is in inverse ratio to the number of parameters. Specially, TPENet-a and TPENet-d have more parameters and lower performance since these deployment locations have a large number of channels, a surge in the number of parameters may lead to network degradation. On the other hand, TPENet-b and TPENet-c have a similar amount of parameters, but the former performs significantly better (47.6% *vs.* 47.0%) TPENet-b has a higher spatial resolution, and STE can achieve better performance.

The number of groups. We also demonstrate the impact of different groups in the TPE module. With $\#G = 1$, and $K_0 = 1$, the limited temporal receptive field makes it difficult to capture long-range temporal context. As observed in Table 1, with $\#G$ increasing, the receptive field of the model expands. Specially, with $\#G = 2$, our method outperforms the TSM baseline (**46.8%** *vs.* 45.6%). The above results show that multi-level temporal modeling is more conducive to fine-grained action reasoning. Furthermore, these lightweight designs achieve large performance improvements without introducing significant computational costs.

Table 1: Study on groups of TPE module. #G: The number of groups

Method	# *G*	Top-1	Top-5	FLOPs
TSM	–	45.2	73.7	32.90G
	1	45.0	73.2	32.92G
TPE	2	46.8	75.6	32.94G
	4	**47.6**	**76.1**	32.97G

Study on the effectiveness of the TPE module. Here, we demonstrate the impact of different sub-module by gradually adding different components. In Table 2, our TP-Module can significantly boost its baseline (20.1% *vs*. 47.6%) and only introduce 0.05G FLOPs. Meanwhile, compared with the SE-Module, our ME achieves a more competitive performance, which shows that the multi-path temporal excitation can facilitate the model to perceive a more discriminative temporal context, thus further releasing the temporal reasoning ability of the proposed network.

Table 2: Study on the effectiveness of the TPE module

Method	Top-1	Top-5	Params	FLOPs
Baseline	20.1	47.4	24.30	32.90
+TP-Module	47.6	76.1	24.34	32.95
+TP-Module+SE	48.4	77.5	24.50	32.97
+TP-Module+ME	**49.4**	**78.3**	24.54	33.10

4.4 Comparison with Other Methods

In Table 3, we compare the proposed method with the existing method to demonstrate its effectiveness. Each section of the table belongs to 2D CNNs-based methods, 3D CNNs-based methods, and the proposed method.

Table 3: Performance and complexity of our method on Sth-Sth V1 compared other methods

Method	Backbone	Frame	FLOPs	Params	Top-1	Top-5
TSN	BNInception	8	16G	10.7M	19.5	–
TRN Two-Stream	BNInception	8	16G × N/A	18.3M	42.0	–
ABM		16	53G	–	47.5	–
TSM		16	65G × 1	24.3M	47.3	77.1
STM	ResNet50	8	33.3G	24.0M	47.5	–
TEINet		8	33G	30.4M	47.4	–
SmallBigNet [33]		8	52G	–	47.0	77.1
NL I3D + GCN [34]	3D ResNet50	32 × 2 clips	303G × 2	62.2M	46.1	76.8
ECO_{EN}Lite	BNIncep + 3D Res18	92	267G	150M	46.4	–
CorrNet-26 [35]	R(2+1)D-26	32	78G	–	47.4	–
TPENet (Ours)	ResNet50	8	33.12G	24.37M	**49.4**	**78.3**
		16	66.24G	24.37M	**50.5**	**80.1**

As observed in Table 3, the TSN baseline failed to achieve good performance on the Sth-Sth V1 due to lack temporal modeling. Our method is obviously superior to the medium and later fusion methods, such as Temporal Relational Networks (TRN), Approximated Bilinear Modules (ABM), and Efficient convolutional networks, which indicates that a single-level of temporal modeling is not enough to represent complex temporal relations. Our TPENet also performs better than the 2D CNNs methods using the same backbone, our methods can significantly enhance the temporal representations by performing multi-temporal aggregation. Furthermore, our method has a more compact topology and significantly outperforms the 3D CNNs-based methods in performance, even better than complex models, such as non-local and graph convolution.

We also evaluate the performance of our method on UCF-101 and HMDB-51 to demonstrate the generalization ability of TPENet on smaller datasets. We fine-tune our TPENet with 16 frames as inputs on these two datasets using model pre-trained on Kinetics-400 and report the mean class accuracy over three splits. From Table 4, compared 3D CNNs-based and 2D CNNs-based methods, our method obtains comparable or better performance, which further demonstrates the effectiveness of TPENet on action reasoning.

Table 4: Comparison with other methods on UCF-101 and HMDB-51 datasets

Method	Backbone	UCF-101	HMDB-51
ECO	BNIncep. + 3D Res18	94.8	72.4
ARTNet	3D ResNet18	94.3	70.9
I3D [2]	Inception V1	95.6	74.8
R(2+1)D	Inception V1	96.8	74.5
TSN	BNInception	91.1	–
TSM	ResNet50	93.5	73.5
STM	ResNet50	96.2	72.2
TEINet	ResNet50	96.7	72.1
TPENet (Ours)	ResNet50	**96.8**	**75.5**

To further explain the effectiveness of TPENet in temporal reasoning, we evaluate its inference ability for complex motion patterns by comparing the keyframe selection results of TPENet and TRN [12].

Specifically, firstly, the test video is divided into 8 segments by the sparse temporal sampling strategy. Then, a frame is randomly sampled from each segment to generate some candidate tuples. Finally, the generated candidate tuple is fed into the network for prediction, and the tuple with the highest test accuracy is selected as the keyframe of the video. The keyframe selection results of the two methods are shown in Fig. 7. Notably, we only show the prediction results of the intermediate keyframes in Fig. 7, since the remaining keyframes selected by the two methods are consistent. In addition, the red text in the figure represents the wrong prediction, the green text represents the correct prediction, and the red box represents the keyframe selected by the two algorithms with the largest gap.

In Fig. 7, TPENet was able to pinpoint the instantaneous critical moment at the edge that was crucial for correct prediction. With the hand gradually moving away from the object, the object can still stand upright on the table and evolves to the next frame, the object starts falling off the table. In addition, the fifth row shows a sample where the prediction fails. The reason for the prediction failure is that the bottle in the picture is right in front of the laptop. From the time the bottle is in front of the laptop until the bottle falls off the table, the model predicts the action as “pretending to put something behind something”. Although there is some ambiguity in this action, thus the failure frequency is much lower.

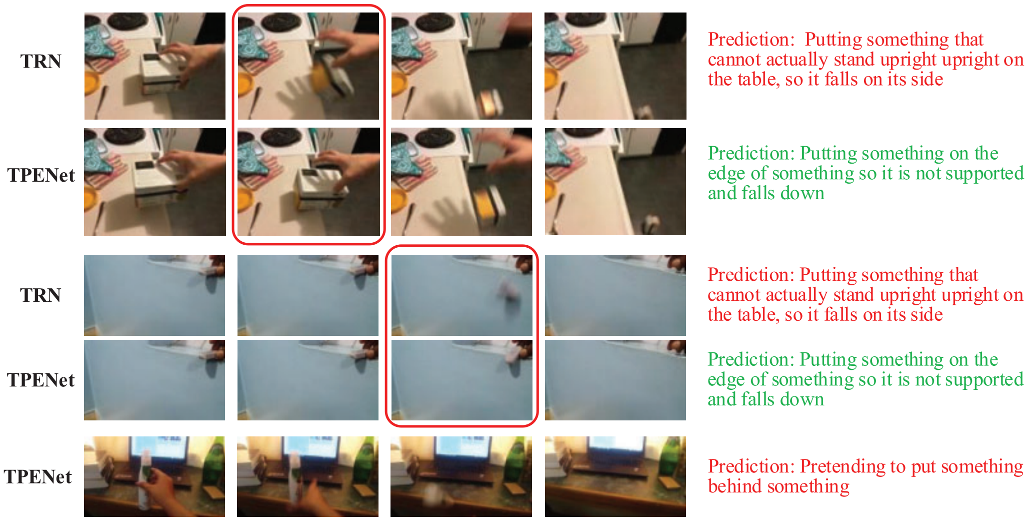
TRN
Prediction: Putting something that cannot actually stand upright upright on the table, so it falls on its side
TPENet
Prediction: Putting something on the edge of something so it is not supported and falls down
TRN
Prediction: Putting something that cannot actually stand upright upright on the table, so it falls on its side
TPENet
Prediction: Putting something on the edge of something so it is not supported and falls down
TPENet
Prediction: Pretending to put something behind something

Figure 7: The action prediction samples of TPENet and TRN

5 Conclusion

Aiming at the limited temporal receptive field in existing methods, a novel temporal pyramid excitation module is proposed in this paper. This module extracts multi-scale temporal context information from video through temporal pyramid convolution, which is beneficial for capturing complex temporal relations and long-term temporal structures. Meanwhile, to overcome the group convolution effect, the Shuffle operation and channel excitation module are adopted in this paper. This combination ensures the interaction between different groups of channels by performing Shuffle and recalibration on the channels of multi-scale temporal features, which is conducive to improving the generalization performance of the model. Extensive ablation experiments and comparative experimental results show that multi-level temporal modeling can improve the performance of fine-grained action recognition compared to modeling approaches with fixed temporal kernels. In addition, the proposed method improves the parallelism and computational efficiency of the model by implementing channel grouping, thus significantly improving the inference speed of the model. In the subsequent visualization research, it can be seen that TPENet can accurately locate the critical moment that plays a decisive role in recognition of actions, which further reflects the advantages of the proposed method in temporal reasoning.

Funding Statement: This research was supported by the research team of Xi'an Traffic Engineering Institute and the Young and middle-aged fund project of Xi'an Traffic Engineering Institute (2022KY-02).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. ICCV*, Santiago, Chile, pp. 4489–4497, 2015.

[2] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the kinetics dataset,” in *Proc. CVPR*, Honolulu, HI, USA, pp. 6299–6308, 2017.

[3] L. Wang, W. Li and L. Van Gool, “Appearance-and-relation networks for video classification,” in *Proc. CVPR*, Salt Lake City, UT, USA, pp. 1430–1439, 2018.

[4] G. Varol, I. Laptev and C. Schmid, “Long-term temporal convolutions for action recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1510–1517, 2017.

[5] G. Bertasius, H. Wang and L. Torresani, “Is space-time attention all you need for video understanding,” *arXiv preprint*. arXiv:2102.05095, 2021.

[6] K. Li, Y. Wang, P. Gao, G. Song, Y. Liu *et al.,* “Uniformer: Unified transformer for efficient spatiotemporal representation learning,” *arXiv preprint*. arXiv:2201.04676, 2022.

[7] D. Neimark, O. Bar, M. Zohar and D. Asselmann, “Video transformer network,” in *Proc. ICCV*, Virtual, Online, Canada, pp. 3163–3172, 2021.

[8] X. Li, Y. Zhang, C. Liu, B. Shuai and J. Tighe, “Vidtr: Video transformer without convolutions,” *arXiv preprint*. arXiv:2104.11746, 2021.

[9] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić *et al.,* “Vivit: A video vision transformer,” in *Proc. ICCV*, Virtual, Online, Canada, pp. 6836–6846, 2021.

[10] N. Hussein, E. Gavves and A. Smeulders, “Timeception for complex action recognition,” in *Proc. CVPR*, Long Beach, CA, USA, pp. 254–263, 2019.

[11] Z. Liu, L. Wang, W. Wu, C. Qian and T. Lu, “Tam: Temporal adaptive module for video recognition,” in *Proc. ICCV*, Virtual, Online, Canada, pp. 13708–13718, 2021.

[12] I. C. Duta, L. Liu, F. Zhu and L. Shao, “Pyramidal convolution: Rethinking convolutional neural networks for visual recognition,” *arXiv preprint*. arXiv:2006.11538, 2020.

[13] B. Zhou, A. Andonian, A. Oliva and A. Torralba, “Temporal relational reasoning in videos,” in *Proc. ECCV*, Munich, Germany, pp. 803–818, 2018.

[14] X. Zhu, C. Xu, L. Hui, C. Lu and D. Tao, “Approximated bilinear modules for temporal modeling,” in *Proc. ICCV*, Seoul, Republic of Korea, pp. 3494–3503, 2019.

[15] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin *et al.,* “Temporal segment networks: Towards good practices for deep action recognition,” in *Proc. ECCV*, Scottsdale, AZ, USA, pp. 20–36, 2016.

[16] J. Lin, C. Gan and S. Han, “TSM: Temporal shift module for efficient video understanding,” in *Proc. ICCV*, Seoul, Republic of Korea, pp. 7083–7093, 2019.

[17] B. Jiang, M. Wang, W. Gan, W. Wu and J. Yan, “STM: Spatiotemporal and motion encoding for action recognition,” in *Proc. ICCV*, Seoul, Republic of Korea, pp. 2000–2009, 2019.

[18] H. Kwon, M. Kim, S. Kwak and M. Cho, “Motionsqueeze: Neural motion feature learning for video understanding,” in *Proc. ECCV*, Glasgow, United Kingdom, pp. 345–362, 2020.

[19] Z. Liu, D. Luo, Y. Wang, L. Wang, Y. Tai *et al.,* “Teinet: Towards an efficient architecture for video recognition,” in *Proc. AAAI*, New York, NY, USA, pp. 11669–11676, 2020.

[20] W. Wu, D. He, T. Lin, F. Li and C. Gan, “Mvfnet: Multi-view fusion network for efficient video recognition,” *arXiv preprint*. arXiv:2012.06977, 2020.

[21] K. Li, X. Li, Y. Wang, W. Jun and Q. Yu, “CT-net: Channel tensorization network for video classification,” *arXiv preprint*. arXiv:2106.01603, 2021.

[22] G. Chen, Y. Zheng, L. Wang and T. Lu, “DCAN: Improving temporal action detection via dual context aggregation,” *arXiv preprint*. arXiv:2112.03612, 2021.

[23] Z. Liu, L. Wang, W. Wu, C. Qian and T. Lu, “Tam: Temporal adaptive module for video recognition,” in *Proc. ICCV*, Virtual, Online, Canada, pp. 13708–13718, 2021.

[24] Z. Qiu, T. Yao and T. Mei, “Learning spatio-temporal representation with pseudo-3D residual networks,” in *Proc. ICCV*, Venice, Italy, pp. 5534–5542, 2017.

[25] K. Hara, H. Kataoka and Y. Satoh, “Can spatiotemporal 3D cnns retrace the history of 2D cnns and imagenet?” in *Proc. CVPR*, Salt Lake City, UT, USA, pp. 6546–6555, 2018.

[26] A. Diba, M. Fayyaz, V. Sharma, M. M. Arzani, R. Yousefzadeh *et al.,* “Spatio-temporal channel correlation networks for action classification,” in *Proc. ECCV*, Munich, Germany, pp. 299–315, 2018.

[27] D. Tran, H. Wang, M. Feiszli and L. Torresani, "Video classification with channel-separated convolutional networks," in *Proc. ICCV*, Seoul, Republic of Korea, pp. 5551–5560, 2019.

[28] Y. Zhou, X. Sun, Z. J. Zha and W. J. Zeng, "MiCT: Mixed 3D/2D convolutional tube for human action recognition," in *Proc. CVPR*, Salt Lake City, UT, USA, pp. 449–458, 2018.

[29] M. Zolfaghari, K. Singh and T. Brox, "ECO: Efficient convolutional network for online video understanding," in *Proc. ECCV*, Munich, Germany, pp. 713–730, 2018.

[30] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal *et al.,* "The "something something" video database for learning and evaluating visual common sense," in *Proc. ICCV*, Venice, Italy, pp. 5843–5851, 2017.

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

[32] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. ICCV*, Barcelona, Spain, pp. 2556–2563, 2011.

[33] X. Li, Y. Wang, Z. Zhou and Y. Qiao, "SmallBigNet: Integrating core and contextual views for video classification," in *Proc. CVPR*, Virtual, Online, USA, pp. 1089–1098, 2020.

[34] X. Wang and A. Gupta, "Videos as space-time region graphs," in *Proc. ECCV*, Munich, Germany, pp. 399–417, 2018.

[35] H. Wang, D. Tran, L. Torresani and M. Feiszli, "Video modeling with correlation networks," in *Proc. CVPR*, Virtual, Online, USA, pp. 352–361, 2020.