http://dx.doi.org/10.62762/TETAI.2024.628205

RESEARCH ARTICLE

LI3D-BiLSTM: A Lightweight Inception-3D Networks with BiLSTM for Video Action Recognition

Fafa Wang[1,2], **Xuebo Jin**[2,*] **and Shenglun Yi**[3]

[1] Beijing iQIYI Technology Co., Ltd., Beijing 100080, China
[2] School of Computer Science and Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China
[3] Department of Information Engineering, University of Padua, 35131 Padua, Italy

Abstract

This paper proposes an improved video action recognition method, primarily consisting of three key components. Firstly, in the data preprocessing stage, we developed multi-temporal scale video frame extraction and multi-spatial scale video cropping techniques to enhance content information and standardize input formats. Secondly, we propose a lightweight Inception-3D networks (LI3D) network structure for spatio-temporal feature extraction and design a soft-association feature aggregation module to improve the recognition accuracy of key actions in videos. Lastly, we employ a bidirectional LSTM network to contextualize the feature sequences extracted by LI3D, enhancing the representation capability for temporal data. To improve the model's robustness and generalization ability, we introduced spatial and temporal scale data augmentation techniques in the preprocessing stage, effectively extracting video key frames and capturing key regional actions. Furthermore, we conducted an in-depth study on spatio-temporal feature extraction methods for video data, effectively extracting spatial and temporal information through the LI3D network and transfer learning. Experimental results demonstrate that the proposed method achieves significant performance improvements in video action recognition tasks, providing new insights and approaches for research in related fields.

Keywords: video action recognition, multi-scale preprocessing, lightweight I3D (LI3D), spatio-temporal feature extraction, bidirectional LSTM.

1 Introduction

With the advent of Internet technology and the rapid proliferation of intelligent devices, people's methods of acquiring information have gradually shifted from text to images and videos. Every moment, a vast array of visual content is either actively or passively received by individuals. Among these forms of media, video occupies the largest proportion.

However, due to the immense volume of data and the rich content encompassed within video information, manually sifting through and analyzing each piece of data is evidently impractical. The advancements in

Submitted: 21 March 2024
Accepted: 02 August 2024
Published: 09 August 2024

Vol. 1, **No.** 1, 2024.
10.62762/TETAI.2024.628205

***Corresponding author:**
✉ Xuebo Jin
jinxuebo@btbu.edu.cn

Citation
Wang, F., Jin, X., & Yi, S. (2024). LI3D-BiLSTM: A Lightweight Inception-3D Networks with BiLSTM for Video Action Recognition. *ICCK Transactions on Emerging Topics in Artificial Intelligence*, 1(1), 58–70.

deep learning and machine learning have introduced novel approaches to implementing algorithms based on massive datasets. Furthermore, the continued evolution of hardware infrastructure has rendered previously unfeasible algorithms achievable through enhanced computational capabilities. Driven by the synergistic progress of computer hardware and deep learning frameworks, the field of computer vision has experienced remarkable growth in recent years. Its applications have extended across diverse sectors, including transportation, entertainment, medicine, industry, and education.

Currently, major video platforms such as YouTube, iQiyi, Tencent Video, and Youku possess an extensive repository of video content. With the burgeoning number of online video users, there has been a significant upsurge in user-generated content being uploaded to these portals. Consequently, the need to automatically understand and categorize video content has become increasingly critical, driving rapid advancement in the field of video action recognition. Evidently, given the sheer volume of online videos, it is impractical to employ manual methods for comparative analysis and filtration. As a result, researchers have turned to machine learning techniques to enhance both the accuracy and efficiency of automated video content understanding.

As we have known, traditional machine learning feature extraction methods are encumbered by several limitations. Firstly, features such as SIFT keypoints exhibit variable dimensionality, resulting in vastly disparate feature dimensions across different videos, thereby precluding direct utilization of SIFT features for video action recognition [1]. Secondly, the immense volume of video resources inevitably leads to the extraction of a plethora of SIFT features, the storage of which consumes substantial computational resources. These challenges pose significant obstacles in industrial applications. In contrast, deep learning methodologies offer the capability to represent video features in uniform dimensions, yielding global feature representations. When coupled with GPU hardware resources, this approach facilitates efficient video feature extraction and understanding, dramatically improving both the speed and precision of video action recognition processes.

The advent of deep learning has provided novel approaches to address these challenges. By representing video features in a unified dimension, deep learning methods can extract global features, significantly enhancing retrieval efficiency and accuracy. Nevertheless, the design of a deep learning model capable of adequately capturing the spatiotemporal information in videos remains an urgent technological challenge.

This research aims to resolve key challenges in video action recognition through innovative deep learning architectures. Specifically, our study will focus on addressing the following three issues:

1. Video Information Optimization: How can we effectively preprocess video data in both temporal and spatial dimensions to maximize the retention of crucial information while minimizing redundancy, thereby providing optimal input for subsequent feature extraction processes?

2. Dynamic Feature Capture: How can we design and improve 3D convolutional networks (LI3D) to more effectively capture spatiotemporal features in videos, particularly enhancing the model's sensitivity and comprehension of dynamic information when processing complex, dynamic scenes?

3. Local Feature Integration: How can we develop an innovative feature aggregation mechanism capable of intelligently identifying and integrating key local features in videos, thereby enhancing the model's understanding and recognition of core video content while avoiding excessive information loss or redundancy?

The resolution of these issues will directly impact the performance and efficiency of video action recognition systems, holding significant implications for intelligent video understanding and large-scale content analysis.

This paper is structured as follows: Section 2 provides an overview of related works in the field. Section 3 details the methodology, covering data augmentation preprocessing, extraction of spatiotemporal features in video based on the LI3D network, and spatiotemporal feature extraction utilizing the LI3D-BiLSTM architecture. The experimental setup is described in Section 4, introducing the dataset used and presenting the results and analysis. Finally, Section 5 concludes the paper, summarizing the key findings and implications of the research.

2 Related Work

Pattern recognition based on video data has made significant progress, with video databases now featuring more diverse actions and larger data volumes.

Video scenarios have evolved from single, simple settings to more complex environments, making video data increasingly representative of real-life situations.

The primary focus of motion pattern recognition algorithms based on video data is how to extract and identify human action features in noisy real-world environments. Andrade-Ambriz et al. [2] achieved human motion state classification by applying temporal convolutional neural networks (TCN) to directly model sequential activity patterns, demonstrating strong performance on sensor-based human activity recognition benchmarks. Karpathy et al. [3] extended CNN-based human action recognition methods by processing stacked image features with fixed windows. Subsequently, Tran et al. [4] applied 3D CNNs to the Sports-1M dataset, enabling the network to effectively capture and recognize temporal features in videos, thus improving human action recognition accuracy. Zha et al. [5] conducted an in-depth comparative study on different strategies using CNNs in video event detection, significantly enhancing both temporal and spatial feature extraction capabilities of CNNs.

The combination of Convolutional Neural Networks (CNN) and Long Short-Term Memory networks (LSTM) has been increasingly successful in recognizing actions in images. This approach uses CNNs to extract feature information from images and LSTMs to process the temporal relationships between consecutive frames, effectively improving action classification results [6]. Donahue et al. [7] further proposed an LSTM-based autoencoder model that maps input temporal image features to fixed-length representations using an encoding LSTM, then uses one or more decoding LSTMs to classify different actions. This model, applied to human action recognition databases UCF-101 and HMDB-51, demonstrated significant improvements in classification accuracy, especially when sample sizes were limited. Zhang et al. [8] proposed ResNeSt, which introduces a split-attention mechanism across feature map groups to capture multi-scale spatial information within a single convolutional block, demonstrating that attention-based feature selection within backbone networks substantially improves representation quality for downstream recognition tasks including video action classification. Li et al. [9] proposed VideoLSTM, which combines convolutional feature extraction with LSTM-based temporal modeling and attention mechanisms to capture both spatial and temporal cues for action recognition in videos.

Many recent motion pattern recognition studies are based on CNNs, LSTMs, and their variants. Wang et al. [10] proposed Temporal Segment Networks (TSN), a video-level framework that uniformly samples multiple short clips from the entire video duration and aggregates segment-level predictions into a video-level consensus, enabling efficient modeling of long-range temporal dynamics for action recognition without prohibitive computational cost. Simonyan and Zisserman [11] proposed two-stream convolutional networks that process spatial appearance features from individual frames and temporal motion information from optical flow through parallel convolutional pathways, demonstrating that explicitly modeling motion cues alongside appearance substantially improves video action recognition accuracy. Diba et al. [12] proposed Temporal 3D ConvNets (T3D), which introduced a temporal transition layer to enable the transfer of pre-trained 2D convolutional weights into 3D networks for video classification, demonstrating that leveraging transfer learning from image-domain models can substantially reduce training data requirements while improving spatiotemporal feature extraction quality—a strategy directly adopted in the design of our LI3D network.

In recent years, video-based motion pattern classification has garnered widespread attention in computer vision. With the advancement of deep learning technologies, researchers have proposed various innovative methods to enhance the accuracy and efficiency of video understanding. The extraction and fusion of spatiotemporal features have become key research directions. Bertasius et al. [13] introduced the TimeSformer model, which pioneered the application of pure Transformer architecture to video understanding tasks, demonstrating the advantages of self-attention mechanisms in capturing long-range spatiotemporal dependencies. However, the high computational complexity of this approach limited its application in resource-constrained environments. To address this issue, Fan et al. [14] proposed the Multiscale Vision Transformer (MViT), which achieves multiscale modeling through hierarchical pooling, maintaining high performance while improving computational efficiency.

In self-supervised learning, Tong et al. [15] extended the concept of masked autoencoders to the video domain, proposing the VideoMAE method, which significantly improved the data efficiency of models

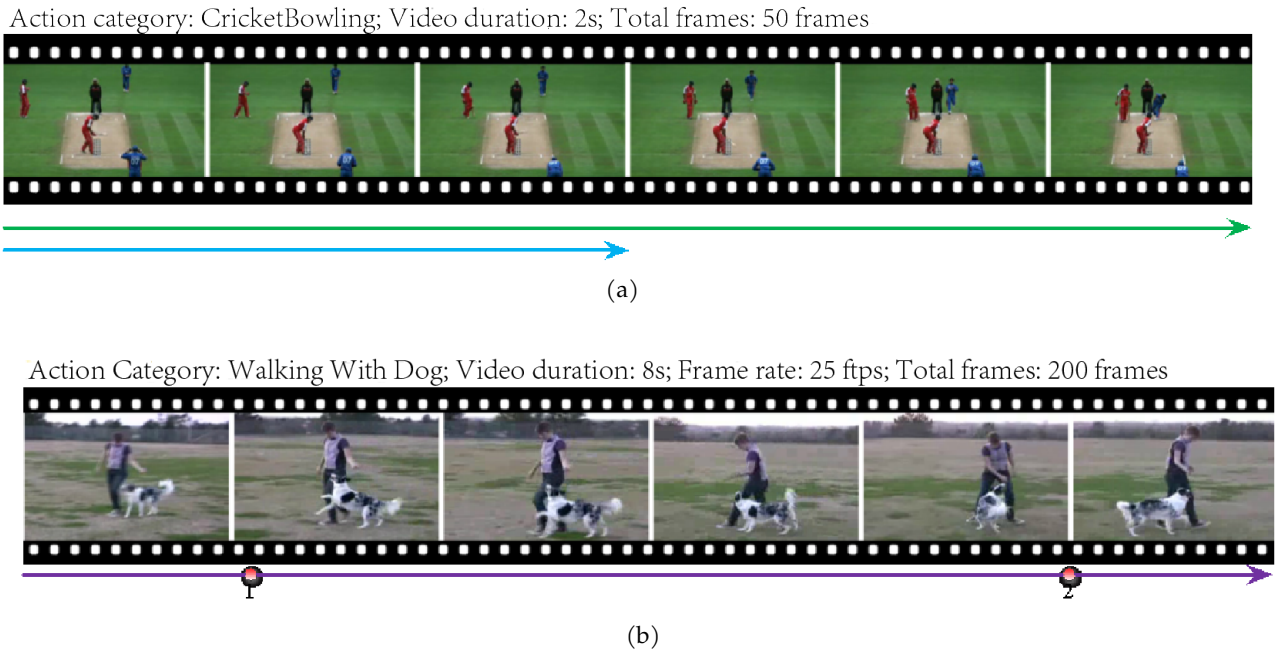

Figure 1. (a) Schematic diagram illustrating the frame extraction strategy for short sequences. (b) Schematic illustration of the frame extraction strategy for extended sequences.

and opened new avenues for utilizing large-scale unlabeled video data. Lin et al. [16] focused on improving the efficiency of video understanding, proposing a temporal shift attention mechanism that cleverly combines the advantages of the Temporal Shift Module (TSM) and self-attention, effectively capturing long-term spatiotemporal dependencies while maintaining low computational complexity. Ryoo et al. [17] introduced the concept of adaptive spatiotemporal tokenization with their TokenLearner, which can dynamically select important regions based on video content, providing greater flexibility in processing different types of videos. Yang et al. [18] proposed a motion-aware masked autoencoder for self-supervised video representation learning, introducing explicit motion signals into the masking strategy to encourage the model to capture dynamic information that is critical for video understanding. Wei et al. [19] proposed a masked feature prediction method, providing a new paradigm for visual pre-training that can learn richer visual representations applicable to various downstream tasks.

These latest studies not only confirm the core status of spatiotemporal feature extraction and fusion in video motion pattern classification but also propose numerous innovative methods to achieve this goal more effectively. They collectively point to several important trends: the widespread application of Transformer architectures in video understanding, the importance of multiscale spatiotemporal modeling, the potential of self-supervised learning in leveraging large-scale unlabeled data, and ongoing efforts to improve computational efficiency while maintaining high performance.

The main contributions of this paper are as follows:

1. To enhance the robustness and generalization ability of the network, we researched data augmentation and applied this technique in the data preprocessing stage, achieving spatial and temporal augmentation of video data. Augmentation operations can effectively extract key frames from videos and capture and represent actions in key areas of the video, greatly enhancing the representational capacity of the data.

2. We studied methods for extracting spatiotemporal features from video data. Video features differ from image features in that they must consider both spatial and temporal dimensions simultaneously. Extracting and modeling only frame-level features would lose the natural connections between consecutive frames in a video. Such isolated features cannot fully represent the content of a video segment. This paper proposes a Lightweight Inception-3D Networks (LI3D) and adopts a transfer learning approach to extract video features. This model

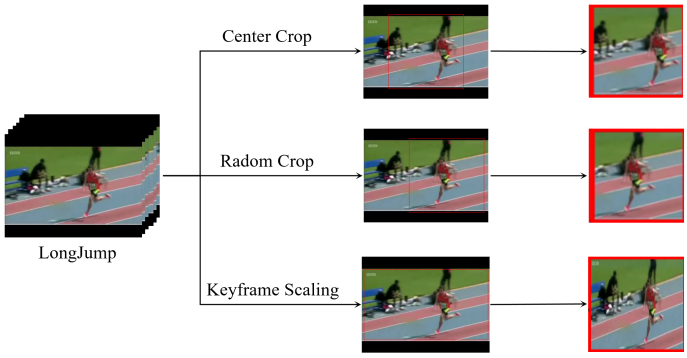

Figure 2. Three data augmentation techniques.

can effectively extract both spatial and temporal information from videos.

3. Dynamic video features differ from static image data. When re-encoding video data, it is crucial to consider the temporal relationships of features. To further enhance the network's ability to recognize temporal features in video data, this paper builds upon the research of Bidirectional Long Short-Term Memory networks (Bi-LSTM). The video feature sequences extracted by the LI3D network are processed through Bi-LSTM for contextual association, enhancing the network's ability to represent temporal data features.

3 Methods

3.1 Data augmentation preprocessing

In the process of preprocessing video data, diverse data augmentation strategies were implemented, taking into account the spatiotemporal characteristics of the footage. Temporally, multi-scale frame extraction was employed, enabling effective key frame selection for videos with varying motion dynamics. This approach ensures that the information fed into the network more accurately represents the video content. Spatially, multi-scale video cropping techniques were applied, utilizing three distinct cropping methods to extract specific content from key frames. This methodology not only enhances the feature diversity of the dataset but also facilitates the acquisition of more comprehensive feature information.

3.1.1 Multi-scale temporal video frame extraction and processing

Given the intricate temporal relationships inherent in video content, utilizing individual frames extracted from a video fails to adequately represent the comprehensive narrative of the entire sequence. To address this limitation, this paper employs a three-dimensional convolutional neural network as the model for extracting deep features from video data, effectively capturing the temporal information within the footage. This section primarily investigates the temporal data preprocessing procedure prior to network input.

To ensure that the algorithm-extracted features sufficiently express the temporal action information within the video, the most crucial element is providing an abundance of temporally correlated consecutive video frames. This information serves as the model's input, facilitating the extraction of features that encapsulate the entire video's content. However, due to the diversity of action categories in videos, some action information is relatively slow and periodic, such as walking or running, which often spans longer image sequences, as illustrated in Figure 1(a). Conversely, other action information is instantaneous and fleeting, such as archery or basketball shooting, covering relatively shorter sequence lengths, as depicted in Figure 1(b). Consequently, during the video data preprocessing phase, this paper proposes an adaptive multi-temporal scale video frame extraction strategy. This approach is capable of extracting key frames that adequately express video content for videos of varying durations and action characteristics.

This sophisticated frame extraction technique enables the original data input into the network to encapsulate a more comprehensive temporal narrative from the video, thereby achieving the objective of enhancing the diversity of training data. It allows for the acquisition of more nuanced temporal features while maintaining the dimensional integrity of the input data. For each video within the dataset, FFmpeg is employed to transmute the footage into a sequence of individual frames, thus ascertaining the total frame count corresponding to each video *Total_frames*, the requisite number of frames for model input is denoted as *Need_frames*. For specific videos, diverse frame extraction strategies are employed.

1) When the total frame count *Total_frames* does not exceed *Need_frames*, as illustrated in Figure 1(a), the requisite number *Times of traversals Total_frames* (denoted by green arrows) and supplementary frames *Left_frames* (indicated by blue arrows) are calculated. Subsequently, the key frames *keyframes* destined for model input are derived from *Total_frames*;

2) When the total number of frames *Total_frames* in a video exceeds a certain threshold *Need_frames*, as illustrated in Figure 1(b), we subtract the required number of key frames *Need_frames* from the total frame

Total_frames count to determine the starting point for random frame extraction *begin_index* (denoted by red point 1). Subsequently, we establish the endpoint *end_index* (marked as red point 2) by adding the desired total frame count to the initial frame position. Within this feasible range, we proceed to extract the key frames *keyframes*. The primary algorithmic process is delineated in Algorithm 1.

Algorithm 1: Multi-temporal scale video frame extraction strategy.

Data: frame_list, Need_frames
Result: keyframes

$Total\_frames \leftarrow len(frame\_list)$;
keyframes = [ ];
if $Total\_frames \leq Need\_frames$ **then**
  $Times = \lfloor \frac{Need\_frames}{Total\_frames} \rfloor$;
  $Left\_frames = Need\_frames \% Total\_frames$;
  for i *in* $range(Times)$ **do**
    keyframes.extend(frame_list);
  end
  keyframes.extend(frame_list[:Left_frames]);
else
  $begin\_index = Random(0, Total\_frames - Need\_frames)$;
  $end\_index = begin\_index + Need\_frames$;
  keyframes = frame_list[$begin\_index$:end_index];
end

3.1.2 Multiscale Spatial Video Trimming and Processing

Temporal data augmentation techniques engender a diverse chronological representation within the training dataset, enabling neural networks to extract more comprehensive temporal features. Similarly, in the spatial domain, employing multi-scale cropping procedures enriches each input frame with a plethora of spatial information, facilitating the extraction of more abundant spatial data and thus achieving spatial data augmentation. The multi-scale video cropping strategies utilized in this study, as illustrated in Figure 2, predominantly incorporate three methodologies: central cropping, random cropping, and key frame scaling.

The specific methodologies are as follows:

(1) Key Frame Scaling: This technique involves directly resizing the original image to the dimensions required by the network. This process enhances the image's smoothness and clarity.

(2) Center Cropping: This method extracts the network-required image size from the central region of each key frame. These key frames can thus acquire more crucial information about the image's central area.

(3) Random Cropping: This approach involves randomly cropping the key frame image to obtain an image that conforms to the network's input dimensions. The specific operational process is as follows:

Assuming the required image width is L_{need}, and the original key frame width is L_{origin}, we can derive a boundary information $margin = (L_{\text{origin}} - L_{\text{need}})/2$. Utilizing this boundary information, we can determine the starting position p_{start} for random cropping, with an interval of $(margin_{\text{min}}, margin_{\text{max}})$. Here, $margin_{\text{min}} = \max(0, 0.8 \times margin)$, $margin_{\text{max}} = \min((1.2 \times margin), L_{\text{origin}} - L_{\text{need}})$, and the cropped image width range is $(p_{\text{start}}, p_{\text{start}} + L_{\text{need}})$.

Through these three spatial data augmentation techniques, we can effectively enhance the network's convergence speed and improve its generalization capabilities.

3.2 Extraction of Spatiotemporal Features in Video Based on LI3D Network

In the process of image content recognition and classification, feature extraction is pivotal in determining the quality of algorithmic identification. This critical issue has given rise to numerous traditional feature description methods, such as SIFT and HOG. In recent years, with the continuous advancement of deep learning, feature descriptors based on Convolutional Neural Networks (CNNs) have been increasingly employed in image content recognition and classification domains. In contrast to conventional feature descriptors, the process of extracting feature descriptors using CNNs is tantamount to training a series of filters (convolutional kernels). These filters are analogous to detection operators in traditional feature extraction methods. The distinguishing factor lies in the fact that detection operators in traditional methods like SIFT and HOG are typically designed by humans, derived from extensive prior knowledge, whereas these filters are autonomously learned through data-driven processes during neural network training.

In the realm of video feature extraction, 3D convolutional networks are utilized to simultaneously

learn temporal and spatial information, yielding descriptors capable of representing both temporal and spatial features concurrently—a feat unattainable through traditional feature extraction methods. However, due to the introduction of convolution operations in the temporal dimension, conventional 3D convolutional neural networks, such as C3D, encounter enormously high computational demands during training. Moreover, the limited availability of video data impedes the provision of superior pre-trained weights. Consequently, C3D's efficiency in extracting video temporal features is suboptimal, and the network's recognition accuracy falls short of excellence.

I3D [21] adeptly addresses these two challenges. Firstly, I3D leverages the Inception-v2 network architecture, which, compared to traditional C3D networks, is deeper and yields more enriched features. Furthermore, I3D can utilize Inception-v2's pre-trained weights from the large-scale ImageNet dataset, significantly reducing the computational requirements for model training and enhancing the network's robustness and generalization capabilities.

The network structure of I3D is derived from Inception-v2 by expanding the dimensions of convolutional kernels, as illustrated in Figure 3. The input video file undergoes 3D convolution and pooling operations for preliminary feature extraction. Subsequently, the extracted video features are further refined through consecutive Inception Module blocks, ensuring that the extracted spatiotemporal features more accurately reflect the genuine motion patterns within the video.

In the realm of deep learning network architectures for extracting spatiotemporal features from videos, I3D boasts two fundamental advantages. Firstly, the I3D model structure incorporates the Inception Module, as illustrated in Figure 4. This module, originally proposed in Inception-v2, enables the network to become both wider and deeper, thus facilitating the acquisition of more diverse spatiotemporal features. Through the Inception Module, the network achieves cross-layer feature fusion, amalgamating these rich feature information streams to enhance the identification of content within images.

Secondly, a notable advantage of the I3D model lies in its utilization of transfer learning principles. By leveraging the ImageNet dataset for pre-training on the Inception Module and subsequently expanding two-dimensional convolution kernels into three

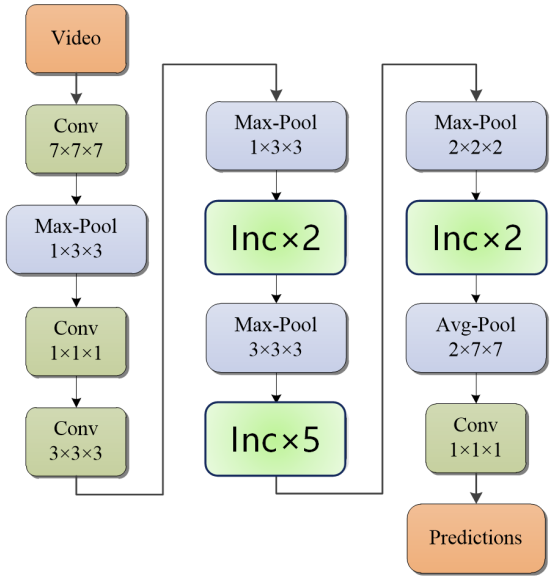

Figure 3. Schematic representation of the I3D network architecture.

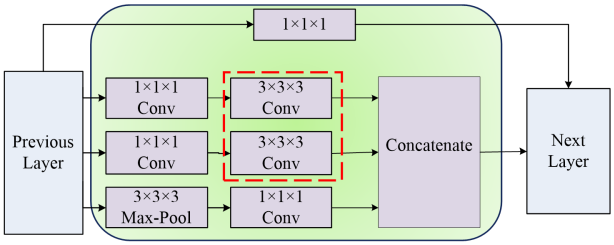

Figure 4. The inception module of I3D architecture.

dimensions, followed by retraining the network using the Kinetics dataset, this approach not only addresses the limitations of insufficient training data but also rectifies the shortcomings of two-dimensional convolutions in capturing temporal features of video content.

However, the presence of 3x3x3 convolution kernels in the I3D network structure still results in substantial computational complexity, thereby reducing the efficiency of feature extraction. To address this issue, we have modified the Inception module of I3D, as illustrated in Figure 5. Drawing inspiration from the structural pattern of Inception-v3 [20], which decomposes large convolutions into asymmetric factorized kernels to reduce computational cost while preserving receptive field coverage, we replace the original $3\times3\times3$ convolutions with a combination of $3\times1\times1$ and $1\times3\times3$ convolution kernels. This modification has reduced our network's parameter count from the initial 12M to 8M, resulting in a more lightweight model. Moreover, the reduction in model parameters enhances the model's generalization capabilities and mitigates the risk of overfitting.

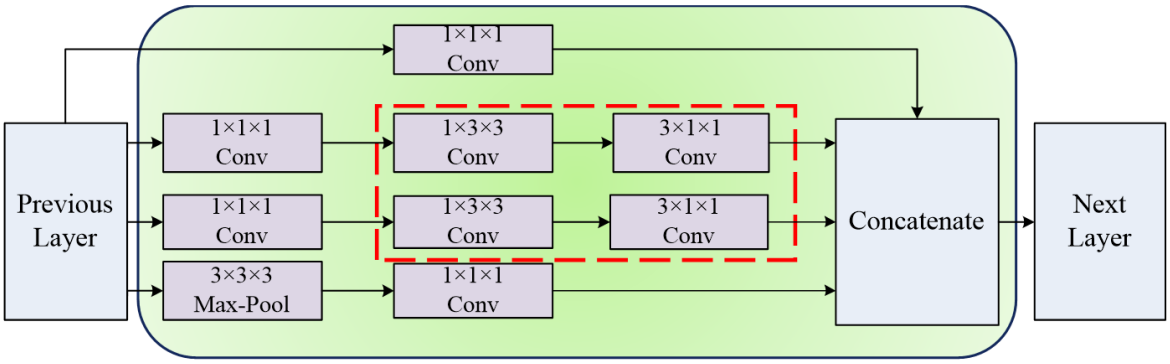

Figure 5. The inception module of LI3D.

Subsequent experiments utilize the improved network model (LI3D) as the baseline for comparative analysis. We extract 1024-dimensional temporal features from the Average pooling layer of the LI3D model. These features are derived from the input network's temporal video features through convolution and pooling operations. The variable τ represents the temporal length, with its dimension determined by the length of the input feature clip, while 1024 corresponds to the spatial feature dimension. The LI3D's extraction of video features yields representations encompassing both temporal and spatial dimensions. By subjecting the LI3D-extracted features to **Bi-LSTM** re-encoding, we fully exploit the contextual relationships within temporal data, further capturing the temporal characteristics of video features and enhancing the representation of video content.

3.3 Soft-Association Feature Aggregation

To further enhance recognition of key actions, we design a soft-association feature aggregation module that dynamically weights the temporal feature clips extracted by LI3D. Given the τ feature clips $\{f_1, f_2, \ldots, f_\tau\}$ extracted from the average pooling layer, the aggregated representation is computed as:

$$F = \sum_{t=1}^{\tau} \alpha_t \cdot f_t, \quad \alpha_t = \frac{\exp(w^\top f_t)}{\sum_{j=1}^{\tau} \exp(w^\top f_j)} \tag{1}$$

where w is a learnable scoring vector and α_t denotes the soft attention weight assigned to each temporal clip, enabling the network to focus preferentially on clips containing discriminative action content.

3.4 Spatiotemporal Feature Extraction Utilizing LI3D-BiLSTM Architecture

To better utilize the features extracted by LI3D and represent the contextual relationships within video content, this chapter employs a Bidirectional Long Short-Term Memory (Bi-LSTM) module, building upon the foundation of recurrent neural networks (LSTM), to enhance the network's recognition capabilities. The Bi-LSTM structure allows the state at any given moment to be determined by both preceding and subsequent inputs, thus incorporating comprehensive past and future contextual information from every point in the input sequence. This approach more accurately reflects the relationships between frames in video content.

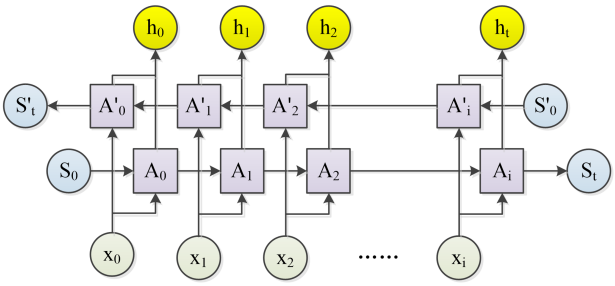

Figure 6. Computational process of bidirectional long short-term memory (Bi-LSTM).

The hidden layer of a Bi-LSTM preserves two values, A and A', which participate in the network's forward and backward computations, respectively. The final output depends on both A and A', as illustrated in Figure 6, which demonstrates that during forward computation, the hidden layer's state output at time t is related to the previous moment, $t-1$. Conversely, during backward computation, the hidden layer's state value at time t is associated with the subsequent moment's state value, $t+1$.

In the feature analysis phase of sequential video data, the utilization of bidirectional recurrent neural networks facilitates a more comprehensive capture of temporal relationships between preceding and subsequent sequences. This approach significantly enhances the representational capabilities of the n-dimensional features obtained during the LI3D

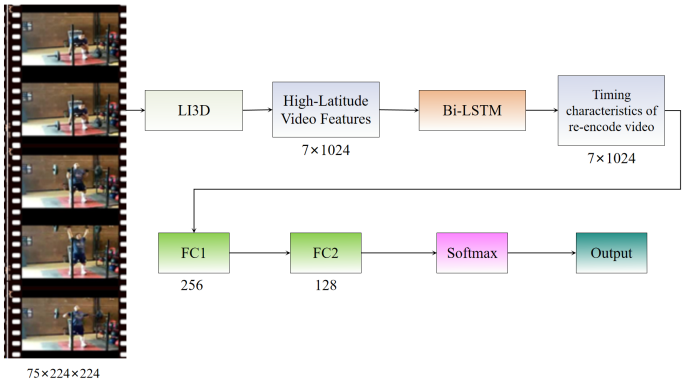

Figure 7. The architectural framework of LI3D-BiLSTM network.

feature extraction stage. The specific feature extraction and analysis process is illustrated in Figure 7.

4 Experiment

4.1 The dataset

The experimental environment utilized an NVIDIA Tesla P40 GPU server. The study employed the open-source UCF101 dataset [22], which comprises 13,320 videos primarily extracted from YouTube, with a cumulative duration of 27 hours, encompassing 101 distinct categories. These categories are classified into five broad groups: human-object interaction, bodily movements, human-human interaction, musical instrument performance, and sports activities. Each category contains 25 videos, with each video featuring four to seven action sequences. This research adhered to the third official partitioning method, dividing the dataset into training and testing subsets. The training subset consists of 9,593 videos, encompassing all five aforementioned activity types. An overview of the UCF101 dataset is illustrated in Figure 8.

4.2 Results and Analysis

Case 1: This study compares the performance of the proposed LI3D network structure with the traditional I3D network structure, evaluating them across multiple dimensions such as network parameters, testing time, accuracy, precision, and recall after 50 iterations. The calculations for accuracy, precision, and recall are as follows:

$$\text{Accuracy} = \frac{\text{Correctly predicted samples}}{\text{Total samples}} \quad (1)$$

$$\text{Precision} = \frac{\text{True positives}}{\text{All predicted positives}} \quad (2)$$

$$\text{Recall} = \frac{\text{True positives}}{\text{All actual positives}} \quad (3)$$

The results of the first experiment, as shown in Table 1, clearly demonstrate that the LI3D network structure effectively reduces the number of network parameters from 12.28M to 8.16M, an overall reduction of approximately 33.5%. This reduction in parameters enhances the network's data fitting capabilities while minimizing the risk of overfitting.

On the UCF101 test set, the accuracy improved from 0.82 to 0.85, precision increased from 0.84 to 0.86, and recall rose from 0.81 to 0.84. Furthermore, the LI3D network's utilization of 3x1x1 and 1x3x3 convolution kernels in place of the I3D network's 3x3x3 kernels significantly reduces computational complexity. This optimization decreases the network's testing time from 2.80s to 1.62s, a reduction of approximately 42.1%, bringing the network's performance closer to real-time efficiency.

Table 1. Performance comparison between I3D and LI3D networks (prior to data augmentation).

Model	Network Parameters	Testing Duration	Accuracy	Precision	Recall
I3D	12.28M	2.80s	0.82	0.85	0.81
LI3D	8.16M	1.62s	0.84	0.86	0.84

Building upon this foundation, we incorporated data augmentation techniques to compare the changes in accuracy, precision, and recall rates between I3D and LI3D models after 50 iterations. The comparative results are presented in Table 2. It is evident that both I3D and LI3D networks exhibited significant improvements in recognition performance following data augmentation. Moreover, LI3D demonstrated a more pronounced enhancement in accuracy, precision, and recall rates compared to I3D. The three metrics increased from 0.85, 0.88, and 0.84 to 0.89, 0.90, and 0.88, respectively. This indicates that after data augmentation, the LI3D network model was able to maintain a high recall rate while simultaneously achieving a high precision for positive predictions. Figures 9 and 10 illustrate the accuracy and loss

Table 2. Performance comparison between I3D and LI3D networks (after data augmentation).

Model	Accuracy	Precision	Recall
I3D	0.85	0.88	0.84
LI3D	0.89	0.90	0.88

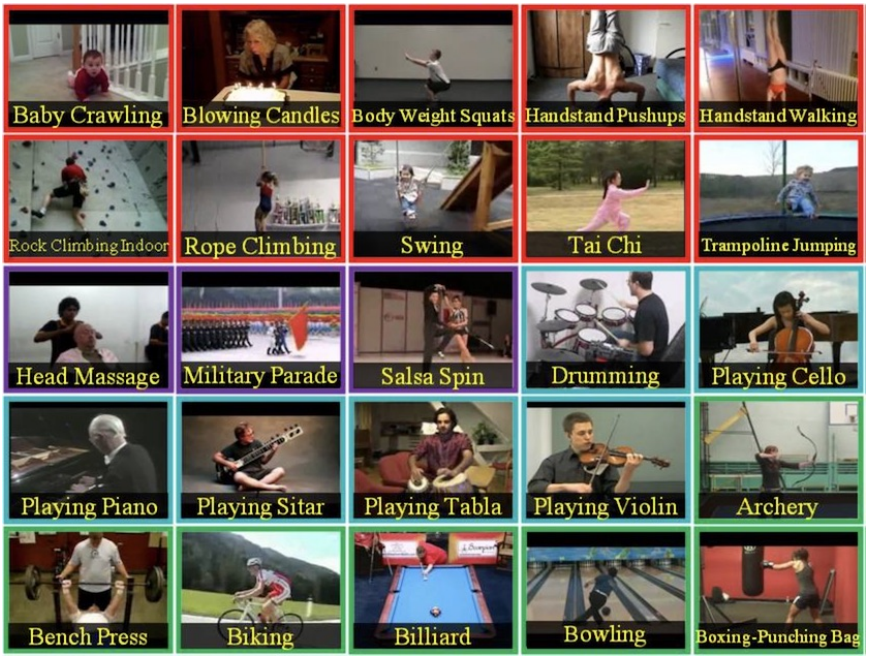

Figure 8. A comprehensive overview of the UCF101 dataset.

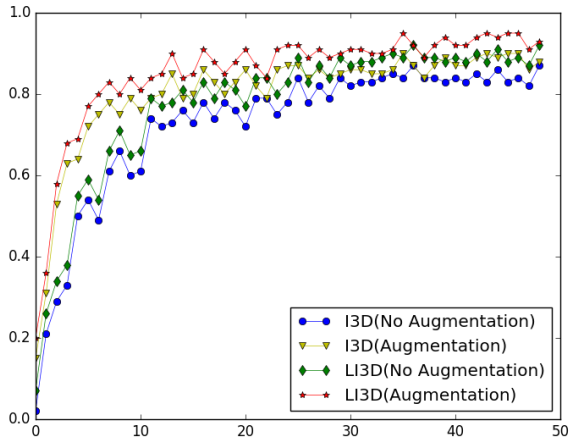

Figure 9. Accuracy curves of I3D and LI3D models before and after data augmentation.

function curves during the training phase for I3D and LI3D networks, both before and after data augmentation, under the same 50-iteration conditions.

The azure and amber curves in Figure 9 illustrate the accuracy trajectories of the conventional I3D model before and after data augmentation, respectively. It is evident from the graph that post-augmentation, the network's accuracy has experienced a substantial enhancement compared to its prior state. However, it is noteworthy that the post-augmentation accuracy of the I3D model merely attains the pre-augmentation level of the LI3D model. In contrast, the LI3D model, under identical conditions of 50 iterations, achieves a remarkably satisfactory accuracy following data augmentation. This observation underscores the efficacy of data augmentation in elevating network recognition performance. Moreover, when coupled with the fine-tuned LI3D model, the improvement in accuracy is particularly pronounced in comparison to the traditional I3D network.

Figure 10 presents a comparative analysis of the loss function curves for I3D and LI3D models over 50 iterations, both with and without data augmentation. The graph elucidates that post-augmentation, the networks exhibit accelerated convergence, enhanced stability, and improved robustness. Notably, the synergy between the LI3D model and data augmentation strategies yields the most rapid convergence among all configurations examined.

Case 2: Building upon the findings from Experiment I, which revealed that the optimal performance configuration consists of data augmentation strategies coupled with the LI3D model, this subsequent investigation aims to juxtapose the efficacy of LI3D, LI3D-LSTM, and LI3D-BiLSTM models under identical conditions of data augmentation and 50 iterations. The primary objective is to elucidate the enhancement effects of temporal feature analysis on the LI3D model. Table 3 presents a comprehensive comparison of accuracy, precision, and recall metrics for LI3D, LI3D-LSTM, and LI3D-BiLSTM models, all subjected to 50 iterations.

Note that the LI3D baseline in Table 3 is evaluated

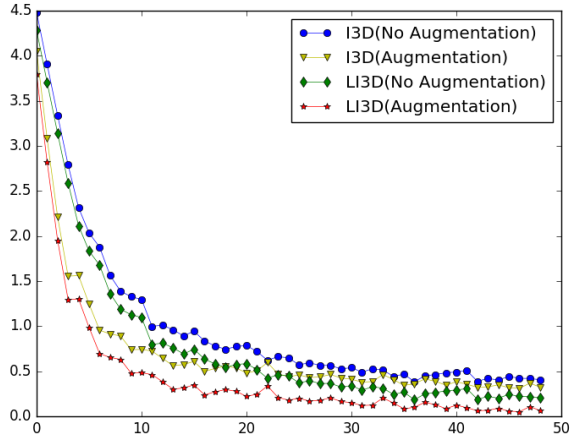

Figure 10. Loss function curves of I3D and LI3D models before and after data augmentation.

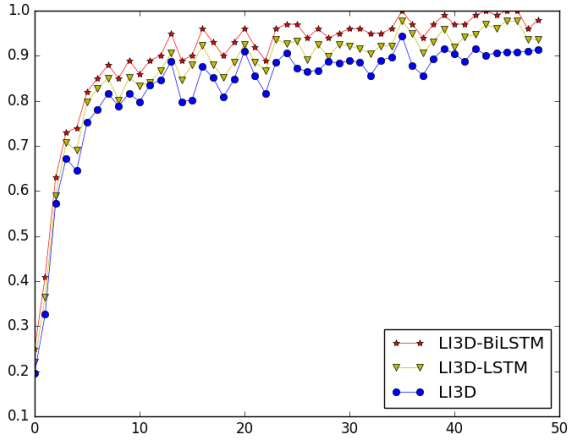

Figure 11. A comparative analysis of training phase accuracy among LI3D, LI3D-LSTM, and LI3D-BiLSTM architectures.

Table 3. Comparative analysis of performance metrics for LI3D, LI3D-LSTM, and LI3D-BiLSTM models (without data augmentation, 50 iterations).

Model	Accuracy	Precision	Recall
LI3D	0.85	0.86	0.84
LI3D-LSTM	0.86	0.89	0.87
LI3D-BiLSTM	0.90	0.91	0.90

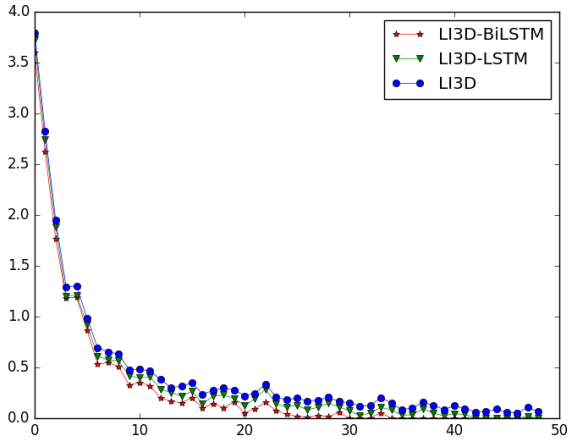

Figure 12. Comparison of loss function for LI3D, LI3D-LSTM, and LI3D-BiLSTM.

without data augmentation to isolate the contribution of temporal re-encoding modules (LSTM and Bi-LSTM) from preprocessing effects, whereas Table 2 reports LI3D performance with data augmentation applied.

From the data presented in the Table 3, it is evident that both LSTM and Bi-LSTM, when applied to temporal features extracted by LI3D, significantly enhance the network's accuracy, precision, and recall rates. This observation underscores the efficacy of LSTM and Bi-LSTM in leveraging temporal relationships within the data, thereby augmenting the features' capacity to represent video content. Moreover, Bi-LSTM demonstrates a superior ability to simultaneously utilize both forward and backward temporal relationships in comparison to LSTM's unidirectional parsing approach, resulting in markedly improved performance.

Figures 11 and 12 illustrate the accuracy and loss function curves for LI3D, LI3D-LSTM, and LI3D-BiLSTM under identical conditions of 50 iterations.

As evident from Figure 11, the LI3D network structure has already achieved a relatively satisfactory level of accuracy. Upon incorporating LSTM for temporal feature re-encoding, the network's accuracy has been further enhanced. Due to Bi-LSTM's ability to utilize features from both time t and t+1 to interpret features at time t, in contrast to LSTM's characteristic of relating features at time t solely to those at time t-1, Bi-LSTM yields feature representations with stronger contextual associations and superior recognition performance. Consequently, LI3D-BiLSTM exhibits higher accuracy compared to LI3D-LSTM.

Figure 12 illustrates that the convergence rate of the LI3D model is considerably slower than that of models incorporating BiLSTM feature interpretation. Networks employing the BiLSTM strategy converge after approximately 30 iterations, indicating that the addition of temporal dimension feature interpretation

strategies results in more stable networks with improved robustness.

5 Conclusion

This research presents an innovative approach to enhance video action recognition, addressing key challenges in the field through a three-pronged strategy. Our method incorporates advanced data preprocessing techniques, a novel lightweight network architecture, and improved temporal feature representation. In the preprocessing stage, we introduced multi-temporal scale video frame extraction and multi-spatial scale video cropping techniques. These methods effectively standardize input formats while preserving crucial content information, significantly improving the model's robustness and generalization capabilities. The implementation of spatial and temporal scale data augmentation further enhanced the extraction of key frames and capture of essential regional actions. Central to our approach is the proposed LI3D architecture, designed for efficient spatio-temporal feature extraction. Coupled with a soft-association feature aggregation module, LI3D demonstrated superior performance in recognizing key actions within videos. The integration of a bidirectional LSTM network further refined the contextual understanding of feature sequences, markedly improving temporal data representation. Experimental results validated the efficacy of our methodology. Comparative analyses of accuracy curves and loss functions revealed that the LI3D model exhibits enhanced stability and faster convergence compared to traditional I3D models. The addition of the BiLSTM module to the LI3D framework yielded substantial improvements in recognition metrics, underlining the effectiveness of our contextual feature association strategy for temporal data representation.

Future research directions include exploring more sophisticated feature aggregation techniques, investigating additional video data augmentation strategies, and extending the application of these methodologies to broader areas within computer vision and multimedia analysis. By continuing to refine and expand upon these techniques, we anticipate further improvements in the accuracy, efficiency, and versatility of video action recognition systems, paving the way for more advanced and reliable applications in real-world scenarios.

Data Availability Statement

The code and data supporting the findings of this study are available from the following link: https://drive.google.com/file/d/1oJMhtkD2SPBO_g5uReY5GZQvltis7-9E/view?usp=sharing.

Funding

This work was supported without any funding.

Conflicts of Interest

Fafa Wang is affiliated with the Beijing iQIYI Technology Co., Ltd., Beijing 100080, China. The authors declare that this affiliation had no influence on the study design, data collection, analysis, interpretation, or the decision to publish, and that no other competing interests exist.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision, 60*(2), 91-110. [CrossRef]

[2] Andrade-Ambriz, Y. A., Ledesma, S., Ibarra-Manzano, M. A., Oros-Flores, M. I., & Almanza-Ojeda, D. L. (2022). Human activity recognition using temporal convolutional neural network architecture. *Expert Systems with Applications*, 191, 116287. [CrossRef]

[3] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 1725-1732). [CrossRef]

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

[5] Zha, S., Luisier, F., Andrews, W., Srivastava, N., & Salakhutdinov, R. (2015). Exploiting image-trained CNN architectures for unconstrained video classification. *arXiv preprint arXiv:1503.04144*. [CrossRef]

[6] Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4694-4702).

[7] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings*

of the IEEE conference on computer vision and pattern recognition (pp. 2625-2634). [CrossRef]

[8] Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., ... & Smola, A. (2022). Resnest: Split-attention networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2736-2746).

[9] Li, Z., Gavrilyuk, K., Gavves, E., Jain, M., & Snoek, C. G. (2018). Videolstm convolves, attends and flows for action recognition. *Computer Vision and Image Understanding, 166*, 41-50. [CrossRef]

[10] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016, September). Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision* (pp. 20-36). Cham: Springer International Publishing. [CrossRef]

[11] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems, 27*.

[12] Diba, A., Fayyaz, M., Sharma, V., Karami, A. H., Arzani, M. M., Yousefzadeh, R., & Van Gool, L. (2017). Temporal 3d convnets: New architecture and transfer learning for video classification. *arXiv preprint arXiv:1711.08200*. [CrossRef]

[13] Bertasius, G., Wang, H., & Torresani, L. (2021, July). Is space-time attention all you need for video understanding?. In *ICML* (Vol. 2, No. 3, p. 4).[CrossRef]

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

[15] Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35, 10078-10093.

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

[17] Ryoo, M., Piergiovanni, A. J., Arnab, A., Dehghani, M., & Angelova, A. (2021). Tokenlearner: Adaptive space-time tokenization for videos. *Advances in neural information processing systems*, 34, 12786-12797.

[18] Yang, H., Huang, D., Wen, B., Wu, J., Yao, H., Jiang, Y., ... & Yuan, Z. (2022). Self-supervised video representation learning with motion-aware masked autoencoders. *arXiv preprint arXiv:2210.04154*. [CrossRef]

[19] Wei, C., Fan, H., Xie, S., Wu, C. Y., Yuille, A., & Feichtenhofer, C. (2022). Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14668-14678). [CrossRef]

[20] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826).

[21] Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4724-4733). [CrossRef]

[22] Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*. [CrossRef]

Fafa Wang a 2019 graduate of Beijing Technology and Business University with a degree in Control Theory and Control Engineering, currently serves as an Algorithm Engineer at iQIYI Technology Co., Ltd. in Beijing. Specializing in computer vision and large-scale model applications, Wang's work encompasses diverse areas including image and video recognition, retrieval, and tracking; NLP-based danmaku comment recognition; large model-based empty shot material recognition; speaker recognition; and video orientation conversion. His background in control engineering, combined with his focus on cutting-edge AI technologies, enables Wang to contribute significantly to various projects at iQIYI, effectively bridging theoretical knowledge with practical applications in artificial intelligence and multimedia processing.(Email: wangfafa@qiyi.com)

Xuebo Jin received the B.S.and M.S. degreesin control theory and control engineering from Jilin University, Changchun, China, in 1994 and 1997, and the Ph.D. degree in control theory and control engineering from the University of Zhejiang, Zhejiang, China, in 2004. From 2009 to 2012, she was an Assistant Professor with Zhejiang Sci-tech University. Since 2012, she has been a Professor with Beijing Technology and Business University, Beijing, China. Her research includes a variety of areas in information fusion, big data analysis, condition estimation, and video tracking.(Email: jinxuebo@btbu.edu.cn)

Shenglun Yi is currently an Assistant Professor in the Department of Information Engineering at the University of Padova, Italy. He received his B.Eng. degree in Automation from Chongqing University, China, in 2016, followed by an M.Sc.Eng. degree in Control Engineering from Beijing Technology and Business University, China, in 2018. In 2022, he completed his Ph.D. in Control Science and Engineering at Beijing Institute of Technology, China. Dr. Yi's research interests encompass a range of topics including robust estimation, information fusion, signal processing, and identification theory. His diverse educational background and current position at a prestigious Italian university demonstrate his international experience and expertise in the field of information engineering and control systems.(Email: shenglun@dei.unipd.it)