Multiscale Fusion of Transformer and Temporal Convolutional Networks for Action Recognition

Rabia Murtaza*

Department of Commerce, University of Central Punjab, Pakistan; tehmasrabia@gmail.com

*Corresponding Author: tehmasrabia@gmail.com

DOI: https://doi.org/10.30211/JIC.202604.007

Submitted: May 22, 2026 Accepted: July 10, 2026

ABSTRACT

Action recognition constitutes an important research direction within computer vision, and it has been widely applied to intelligent monitoring, motion analysis, virtual reality and other practical fields. Most current approaches fail to effectively extract local temporal characteristics while modeling global context relationships, which restricts their capability when handling complicated long-duration action samples. Targeting this challenge, we design an innovative multi-module integrated network named WTCNet (Whale-Transformer-TCN Network). It leverages temporal convolutional networks to extract fine-grained temporal features, adopts Transformer to establish long-range global correlations, and applies the whale optimization algorithm to tune hyperparameters, so as to boost model performance and running efficiency. Comparative experiments are conducted on UCF-101 and Kinetics-400, two mainstream public datasets. The proposed method obtains 92.5% Top-1 accuracy and 98.3% Top-5 accuracy on UCF-101, and gains 82.1% Top-1 accuracy and 94.8% Top-5 accuracy on Kinetics-400, surpassing mainstream baseline algorithms. Meanwhile, our network shows obvious advantages in inference and training speed, verifying its high efficiency and application potential. This work delivers an effective scheme for challenging action recognition tasks, and also inspires follow-up research on multi-module combination and long-sequence learning. The proposed WTCNet can serve as a useful reference for both theoretical exploration and real-world deployment in action recognition.

Keywords: Deep learning, Action recognition, Human pose estimation, Transformer, Recurrent vision transformer, Real-time action

1. Introduction

Video-based motion classification is a core computer vision task aimed at detecting and categorizing human and object motions in video frames. Driven by the booming development of intelligent monitoring, sports data analysis, virtual and augmented reality technologies, this visual analysis technique has been extensively deployed in diverse practical scenarios[1]. For instance, in smart surveillance, action recognition can automatically detect and analyze potential abnormal

behaviors; in sports analysis, it aids in evaluating athletes' movements and techniques; and in VR and AR, action recognition forms the basis for interaction between users and virtual environments[2,3]. The core challenge of action recognition lies in extracting meaningful information from temporal data, especially capturing dynamic changes in motion and patterns of behavior.

Conventional action recognition algorithms predominantly adopt manually designed descriptors including optical flow, HOG and MHI for feature extraction[4]. Though these techniques perform well on simple recognition tasks, they exhibit poor adaptability to high-dimensional data, long video sequences and intricate human actions with the expansion of dataset scale and task difficulty[5]. In recent years, deep learning technologies represented by CNNs and temporally superior RNNs have been rapidly developed, becoming the mainstream technical scheme for video action recognition research[6,7,8]. Nevertheless, current deep learning models usually focus on single feature learning, extracting either local temporal characteristics or global context information alone, while failing to achieve effective fusion of the two features. Such inherent defects lead to unstable recognition results in complex and dynamic action scenarios[9].

To solve the above feature modeling defects, this study constructs an integrated model based on Temporal Convolutional Networks (TCNs) and Transformer with a multi-scale fusion mechanism. TCN modules are responsible for extracting local temporal correlation features and learning short-term action rules, while Transformer structures capture overall context information and fit long-distance sequence dependencies[10]. Furthermore, the Whale Optimization Algorithm (WOA) is introduced to optimize model architecture and hyperparameters. With its outstanding global optimization ability, WOA can search for optimal parameter combinations in complex spaces, which effectively improves model recognition precision and accelerates training convergence[11]. Integrating the strengths of the three modules, the proposed method achieves reliable and efficient action recognition for complex dynamic scenes, striking a balance between high identification accuracy and real-time inference capability to satisfy the practical application requirements of accuracy and efficiency.

The main contributions of this paper are summarized as follows:

1. A multi-scale fusion model that integrates TCNs and Transformers is proposed, fully leveraging the strengths of TCNs in local temporal modeling and Transformers in global context capture.

2. The Whale Optimization Algorithm (WOA) is introduced for hyperparameter optimization, significantly improving training efficiency and accuracy.

3. Extensive experiments on multiple diverse datasets demonstrate the strong generalization ability of the proposed WTCNet model, showcasing its excellent adaptability and robustness across tasks, providing a solution with high generalization capability for action recognition.

The remainder of this paper is organized as follows. Section 2 reviews previous relevant studies and concludes existing research outcomes. Section 3 details the proposed WTCNet framework, clarifying the design principles of TCN, Transformer and WOA, and illustrating the multi-scale fusion mechanism for their collaborative optimization. Section 4 describes the experimental setup and

analyzes the results to verify model validity. Finally, Section 5 summarizes the core findings of this work and prospects future research directions in model optimization and multimodal fusion.

2. Literature Review

2.1 Traditional Methods of Action Recognition

Early action recognition methods primarily relied on handcrafted feature extraction techniques, which typically involved capturing local motion information from videos or images to recognize actions. One of the representative methods is optical flow, which estimates object movement by analyzing pixel changes between consecutive frames, capturing local motion patterns[12]. Optical flow has achieved some success in simpler action recognition tasks, especially in scenarios with a static background and fixed scenes. However, in dynamic environments, complex backgrounds, or under varying lighting conditions, optical flow methods are often susceptible to noise and external interference, resulting in reduced recognition accuracy and weaker robustness[13].

To overcome the limitations of optical flow, researchers have proposed several feature extraction methods based on local descriptors. For instance, the HOG is a classic image descriptor that calculates the gradient information of local regions within an image to describe shape and texture features. While HOG has been effective in static images or simpler action tasks, it often struggles in dynamic scenes, particularly with large-scale motion changes[14]. To address this issue, MHI, an extension of optical flow, focus on capturing object motion patterns in videos, particularly suited for representing the temporal characteristics of actions. MHI describes the motion information from multiple video frames and can effectively capture temporal variations in dynamic scenes. However, challenges remain when dealing with complex backgrounds and local motion occlusion[15]. Traditional methods face difficulties in automatically learning high-level features from data, often requiring human intervention to design specific feature extraction algorithms. Moreover, these methods typically fail to capture global contextual information or long-range dependencies in video data, which results in degraded performance when handling high-dimensional, long-duration sequences[16]. As datasets grow larger and tasks become more complex, the limitations of these methods become increasingly apparent, especially in modern action recognition tasks that involve complex scenes, diverse actions, and long-term dependencies.

Additionally, deep learning techniques are effective at capturing spatiotemporal features in videos and modeling global dependencies in long-duration data, addressing the shortcomings of traditional methods in long-term sequence and complex dynamic modeling[17,18]. However, even with these advancements, current deep learning models still face challenges in integrating spatiotemporal features and capturing long-term dependencies[19,20]. Therefore, the critical research question remains: how to combine local temporal information with global contextual data to further improve the recognition capability and robustness of models. In this context, this paper proposes the WTCNet model, which integrates TCNs and Transformers to fully leverage the advantages of both local temporal features and global contextual information, enhancing performance in complex action recognition tasks.

2.2 Application of Deep Learning in Action Recognition

Driven by the booming advancement of deep learning, especially the mature application of CNNs and RNNs in spatial feature extraction and time-series modeling, deep learning-based techniques have become the mainstream solution for action recognition research[21]. Unlike conventional manual feature extraction algorithms, deep learning models can autonomously mine effective feature representations from original video data, removing the reliance on artificially designed feature modules[22,23]. Such unique superiority makes these methods well-suited for processing massive and complex datasets, with prominent capabilities in learning video spatiotemporal characteristics and establishing long-distance sequence associations. As core deep learning architectures, CNNs are extensively adopted to extract spatial features from video frames. They acquire local spatial attributes including texture and shape via convolution calculations on continuous frame images[24,25]. The emerging 3D-CNN technology expands 2D convolution to three-dimensional computation, integrating spatial X-Y dimension and temporal T dimension to jointly extract video spatiotemporal features[26]. Owing to its powerful spatiotemporal fusion ability, 3D-CNN has been widely applied in various action recognition frameworks and achieved promising results[27]. Nevertheless, it suffers from obvious defects in long-sequence processing: model training difficulty rises sharply, and long-term temporal dependency modeling cannot be guaranteed[28]. To remedy this flaw, RNNs and its optimized variants including LSTM and GRU are proposed. These improved cyclic structures effectively solve the gradient disappearance and explosion problems of original RNNs in long-sequence training, and perform excellently in learning long-term temporal dynamic changes and continuous action rules in videos[25].

Although current deep learning methods have greatly promoted the development of action recognition, multiple bottlenecks still exist in practical applications. These data-driven models heavily rely on sufficient labeled samples and high-performance computing resources, where data quality and computational efficiency largely restrict model training effects in complex scenarios. In addition, mainstream CNN and RNN architectures only focus on local spatial or temporal feature learning, showing inherent deficiencies in global context perception and long-range dependency capture[29]. For complex scene videos and ultra-long action sequences, existing network structures fail to fully mine global effective information, which severely restricts the overall recognition performance. Targeting the above limitations, this study integrates TCN and Transformer architectures. This hybrid design makes up for the defects of traditional CNNs and RNNs in global feature modeling and long-sequence learning, providing an innovative and effective technical scheme for video action recognition tasks.

3. Methodology

3.1 Overview of the Overall Model Architecture

This work develops a novel WTCNet (Whale-Transformer-TCN Network) to strengthen the temporal feature learning and global correlation modeling capacity for complex video action recognition. As depicted in Figure 1, the overall framework integrates three core units: TCN,

Transformer and WOA, which cooperate synergistically to realize high-precision and high-efficiency action recognition. The TCN unit undertakes local temporal feature extraction for video sequences. Equipped with causal and dilated convolution operations as well as residual connection structures, it efficiently mines short-term frame correlation features, stabilizes model training and improves convergence speed. The causal convolution design strictly avoids future information leakage, complies with the time-series causality of video data, and outputs fine-grained short-term motion feature sequences. On this basis, the Transformer unit takes the local temporal features generated by TCN as input and applies the multi-head self-attention mechanism to mine long-range global dependencies. Different from the local feature learning of TCN, Transformer focuses on capturing overall action context and long-sequence dynamic rules, greatly improving the model's comprehension ability for complex and long-duration video actions.

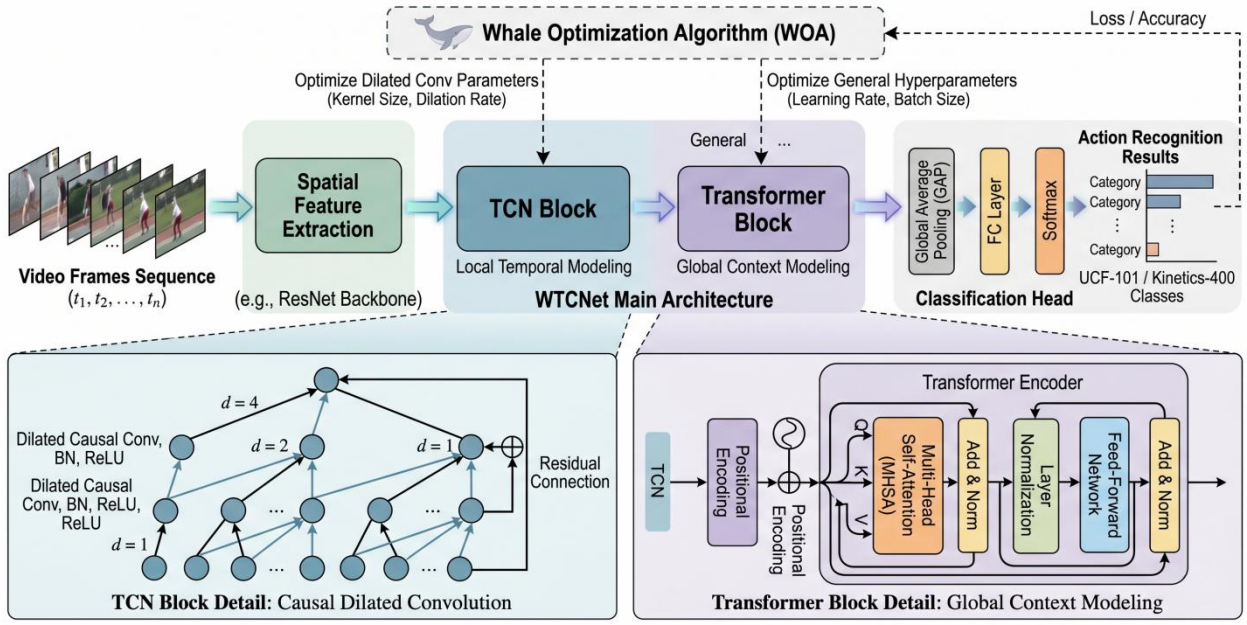

Figure 1. Schematic of the WTCNet model architecture: multi-scale fusion of TCN, transformer, and WOA.

The WOA heuristic optimization algorithm is introduced to dynamically tune the core hyperparameters of the integrated network. Inspired by whale predation behaviors, this optimization algorithm conducts global traversal search to adaptively optimize critical parameters including learning rate, batch size and convolution kernel dimension. This optimization strategy effectively elevates the recognition accuracy, accelerates model training, and enhances the generalization and environmental robustness of the network. The three functional units form a complete and complementary working mechanism: TCN provides reliable local temporal feature basis, Transformer supplements global long-sequence modeling capability, and WOA optimizes network configuration to reach optimal training performance. Benefiting from this collaborative multi-module design, WTCNet possesses outstanding adaptability and stability, achieving superior performance in

processing diverse action types and complex long temporal sequence recognition tasks.

3.2 Temporal Convolutional Network (TCN) Module

The TCN is a core feature extraction unit of the proposed WTCNet, mainly tasked with mining local temporal patterns and short-term frame dependencies in video data. To strengthen the network's temporal modeling capability, this module adopts integrated causal and dilated convolution structures, which effectively optimize the long-sequence learning performance of the model[10,30,31]. The TCN builds a multi-layer temporal processing architecture; each convolutional layer hierarchically refines input features and extracts high-order temporal semantic information for subsequent feature analysis.

Causal convolution is adopted to guarantee rigorous temporal causality during feature extraction. This mechanism restricts convolution calculations to current and historical time steps only, completely eliminating future data leakage in time-series modeling. Its calculation formula is defined as:

$$y_t = \sum_{i=0}^{k-1} x_{t-1} \cdot w_i \quad \text{[Formular 1]}$$

where y_t denotes the output feature at time step t, x_{t-1} refers to the valid input sequence of current and previous moments, w_i represents convolution kernel weights, and k stands for kernel size.

To enlarge the temporal receptive field with negligible computational overhead, dilated convolution is embedded into the TCN structure. It expands feature coverage via adjustable dilation factors, enabling the network to capture long-range temporal correlations efficiently. The corresponding formula is shown below:

$$y_t = \sum_{i=0}^{k-1} x_{t-d \cdot i} \cdot w_i \quad \text{[Formular 2]}$$

where d is the dilation coefficient, and $x_{t-d \cdot i}$ indicates the interval-sampled input features. This design endows each convolutional layer with wide-range temporal perception while maintaining low computing consumption, optimizing the model's long-dependency modeling performance.

Moreover, residual connections are added to convolutional layers to solve the gradient vanishing problem inherent in deep network stacking. The specific calculation is as follows:

$$y_t = F(x_i, \{W_i\}) + x_t \quad \text{[Formular 3]}$$

where $F(x_i, \{W_i\})$ is the learned feature after convolution transformation, x_t is the original input feature, and y_t is the final output of the residual structure. Residual connections retain original feature information during layer transmission, accelerates model convergence, and stabilizes the training process of deep networks.

The stacked dilated convolution layers endow TCN with powerful deep temporal modeling capacity. Each layer extracts fine-grained local temporal features and gradually synthesizes high-level action rules. To enrich feature representation diversity, a multi-scale convolution strategy is adopted. By configuring diverse kernel sizes and dilation coefficients, the module captures multi-scale temporal dependencies to enhance feature generalization. The multi-scale convolution output is formulated as:

$$y_t = \sum_{J=1}^{N} \sum_{I=0}^{K-1} x_{t-d_j \cdot i} \cdot w_{j,i} \quad \text{[Formular 4]}$$

where d_j is the dilation factor of the j-th scale, N is the total number of multi-scale branches,

and $w_{j,i}$ is the matching kernel weight of each scale.

The multi-scale temporal features extracted by the TCN module are finally transmitted to the Transformer module. As a front-end feature extractor, TCN provides refined local temporal details, compensates for the insufficient short-term feature perception of global modeling networks, and forms a complementary cooperation mechanism with Transformer. This paired design significantly improves the overall modeling performance of WTCNet for complex and long video action sequences.

3.3 Transformer Module

As a core component of WTCNet, the Transformer[16,32] focuses on modeling global temporal correlations and implicit associations within long video sequences.

The input sequence $\mathrm{X} = [\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_\mathrm{n}]$ is first fed into the self-attention unit and mapped into Query, Key and Value matrices via learnable weights:

$$Q = XW_Q, K = XW_K, V = XW_V \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\text{[Formular 5]}$$

where $\mathrm{W_Q}$, $\mathrm{W_K}$, and $\mathrm{W_V}$ denote trainable parameter matrices. We compute the correlation between all positions by the dot product of Q and K^T, and scale the result by the dimension d_k of Key vectors to avoid numerical overflow:

$$A = \frac{QK^T}{\sqrt{d_k}} \quad \text{[Formular 6]}$$

After that, Softmax normalizes the similarity matrix into valid attention weights:

$$A_{norm} = softmax(A) \quad \text{[Formular 7]}$$

The final attention output is obtained by weighting the Value matrix with normalized coefficients:

$$O = A_{norm}V \quad \text{[Formular 8]}$$

This workflow lets each position fuse information across the whole sequence to excavate global temporal links.

We adopt multi-head attention to enrich feature representation. Multiple independent attention heads run in parallel, and their outputs are concatenated and mapped with a unified weight matrix:

$$MultiHead(Q,K,V) = Concat(head_1, head_2, \cdots, head_h)W^O \quad \text{[Formular 9]}$$

Each head holds exclusive trainable weights to extract diverse temporal characteristics. Combined with positional encoding to retain sequence order, this module works collaboratively with TCN, greatly boosting the model's capability for long-sequence and complex action recognition.

3.4 Whale Optimization Algorithm (WOA) Module

WOA serves as the optimization unit of WTCNet, dedicated to tuning network hyperparameters and structural settings. This swarm intelligence algorithm mimics whale predation to search for global optima in high-dimensional spaces, lifting model precision and training efficiency[33,34]. Its workflow covers four core phases: group initialization, fitness assessment, position iteration and convergence judgment, which jointly drive the network to reach its best state.

The algorithm first initializes the coordinate of each whale agent. Every vector stands for a group of candidate hyperparameters. For a total of N agents, the i-th individual is defined as:

$$X_i = [x_{i1}, x_{i2}, \cdots, x_{id}], i = 1,2,\cdots,N \quad \text{[Formular 10]}$$

where x_{ij} refers to the j-dimensional value corresponding to one hyperparameter. Next, each agent is scored via the fitness function, which takes model accuracy or loss as evaluation metrics:

$$f(X_i) = Accuracy \ or \ f(X_i) = Loss \quad \text{[Formular 11]}$$

Agents then update their coordinates toward the current optimum following the rules below:

$$X_i^{t+1} = X_i^t + A \cdot D \quad \text{[Formular 12]}$$

$$D = |C \cdot X_i^t - X^*| \quad \text{[Formular 13]}$$

Here X_i^{t+1} and X_i^t are positions at two consecutive iterations, A controls search scope, X^* denotes the optimal agent, and random coefficient C ranges from 0 to 2. To escape local optima, partial agents execute random exploration to enrich population diversity:

$$X_i^{t+1} = X_i^t + \alpha \cdot (X_r - X_i^t) \quad \text{[Formular 14]}$$

where α is the exploration factor and X_r represents a randomly picked agent.

Distinct from gradient-based optimization methods, WOA requires no gradient calculation, so it is free from gradient anomalies in deep network training. Featuring strong global exploration and low computation overhead, it efficiently traverses complex hyperparameter spaces. In WTCNet, WOA coordinates with TCN and Transformer to adjust network configurations adaptively, making the model well-suited for diverse action recognition tasks and maintaining steady performance across different datasets.

4. Experiment

4.1 Datasets

This paper adopts two mainstream action recognition benchmarks, UCF-101 and Kinetics-400, which feature abundant video samples and diverse complex action categories, to comprehensively validate the generalization ability and practical superiority of the proposed WTCNet model.

UCF-101[35] is a video dataset containing 101 action categories, commonly used to assess the performance of action recognition algorithms. The dataset includes 13,320 video samples, with each video typically lasting between a few seconds and a few dozen seconds. It covers a wide range of action types, from sports to everyday activities. Each video is labeled with a specific action category, and videos within the same category exhibit considerable variations in shooting environments, backgrounds, and human poses. The diversity within UCF-101 poses particular challenges in testing the robustness of WTCNet in complex backgrounds.

The Kinetics-400 dataset[36], on the other hand, is a larger and more complex dataset containing 400 action categories and over 300,000 video samples. Compared to UCF-101, Kinetics-400 not only features a larger number of categories but also presents a greater diversity of video samples within each category, encompassing a wide range of scenarios from daily life to professional sports. The dataset is designed to push the performance of models on large-scale, long-duration video data, especially in the context of video classification tasks. Kinetics-400 videos tend to be longer, typically ranging from tens of seconds to several minutes, with strong continuity and temporal dependencies in the actions.

Both UCF-101 and Kinetics-400 are benchmark datasets in the field of action recognition. They

not only offer a rich variety of action categories and complex background noise but also feature different video durations and varying types of temporal dependencies. The significant differences in video content, background diversity, and action representations make these datasets crucial tools for evaluating the adaptability and robustness of deep learning models across various scenes and tasks. By conducting experiments on these two datasets, we are able to thoroughly examine the performance of WTCNet on action recognition tasks of varying scales and complexities, gaining deeper insights into its strengths and weaknesses in real-world applications. Furthermore, this helps advance the development of action recognition technologies in complex environments.

4.2 Experimental Setup and Configuration

This section details the experimental environment and parameter settings adopted in this work to verify the overall performance of the proposed WTCNet on different datasets. All experiments were implemented under unified hardware and software conditions to guarantee fair and credible comparative analysis. The specific experimental configurations are listed in Table 1, covering operational equipment, development framework, and core training hyperparameters.

Table 1. Detailed experimental environment and parameter settings.

Experimental Condition	Specific Configurations
Computation Device	Four NVIDIA A100 GPUs with 128 GB total memory
System Environment	Ubuntu 20.04 Operating System
Development Platform	PyTorch 1.12.0 Deep Learning Development Library
Parameter Updater	Adam optimizer ($\beta_1 = 0.9,\ \beta_2 = 0.999$)
Initial Learning Rate	1e-4, equipped with adaptive decay strategy
Training Batch Size	64 for UCF-101 dataset; 32 for Kinetics-400 dataset
Training Iterations	50 epochs for UCF-101 dataset; 100 epochs for Kinetics-400 dataset
Optimization Metric	Cross-Entropy Loss Function
Sample Enhancement	Random crop, horizontal flip, color distortion and other operations

Source: By authors.

The experiments were conducted on a high-performance computing platform equipped with four NVIDIA A100 GPUs and 128 GB of large memory, which provides sufficient computing power and data caching space to support efficient training of the complex network and large-scale video datasets. The entire experiment ran on the Ubuntu 20.04 system, with PyTorch 1.12.0 selected as the development framework for its flexible dynamic graph mechanism, which greatly facilitates model construction, debugging and iterative training. During the training phase, the Adam optimizer was used for parameter updating with a preset initial learning rate and decay strategy. Differentiated batch

sizes and training epochs were configured for UCF-101 and Kinetics-400 datasets to adapt to their data scales. Meanwhile, common data augmentation methods including random cropping, horizontal flipping and color jittering were adopted to enrich training samples, effectively enhancing the generalization and anti-interference ability of the WTCNet model.

4.3 Evaluation Metrics

To fully and objectively verify the comprehensive performance of the proposed WTCNet in complex action recognition scenarios, this work adopts five mainstream evaluation indicators for multi-dimensional quantitative analysis, covering recognition accuracy, classification balance, training cost and practical inference efficiency. These multi-angle metrics can effectively reflect the model's comprehensive competitiveness in both recognition capability and application practicability.

In this paper, Top-1 and Top-5 accuracy are adopted to evaluate the overall classification precision of the model. Top-1 accuracy calculates the proportion of fully correct predictions, while Top-5 accuracy judges whether the real label falls within the model's top five prediction results, as defined below:

$$Top-1=\frac{\sum_{i=1}^{N}\mathbb{I}(y_i=\hat{y}_i)}{N} \quad \text{[Formular 15]}$$

$$Top-5=\frac{\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i^{(1)}=y_i\cup\hat{y}_i^{(2)}=y_i\cup\cdots\cup\hat{y}_i^{(5)}=y_i)}{N} \quad \text{[Formular 16]}$$

where N represents the total number of test samples, y_i denotes the ground-truth label, \$\hat{y}_i\$ refers to the predicted result, and $I(\cdot)$ is the indicator function.

$$F1\ Score=2\cdot\frac{Precision\cdot Recall}{Precision+Recall} \quad \text{[Formular 17]}$$

$$Training\ Time=T_{end}-T_{start} \quad \text{[Formular 18]}$$

$$Inference\ Speed=\frac{N}{T_{inference}} \quad \text{[Formular 19]}$$

where T_{start} and T_{end} are the start and end time of model training, N is the total number of test samples, and $T_{inference}$ denotes the total inference time for all test data. Combined with the above indicators, this study achieves a comprehensive evaluation of WTCNet's recognition accuracy and operational efficiency.

4.4 Result

Figure 2 plots the training and validation loss curves of WTCNet on UCF-101 and Kinetics-400. For UCF-101, loss values drop sharply at the initial training phase and tend to stabilize after the 10th epoch. Minor oscillations appear in the late-stage validation loss, a typical sign of slight overfitting, yet the overall value stays under 0.2, which proves reliable generalization. On Kinetics-400, the loss variation follows a similar general trend but with gentler fluctuations. Though this dataset features larger data volume and more intricate action types, the model converges steadily: loss declines rapidly in early iterations and maintains a low stable level throughout the whole training process.

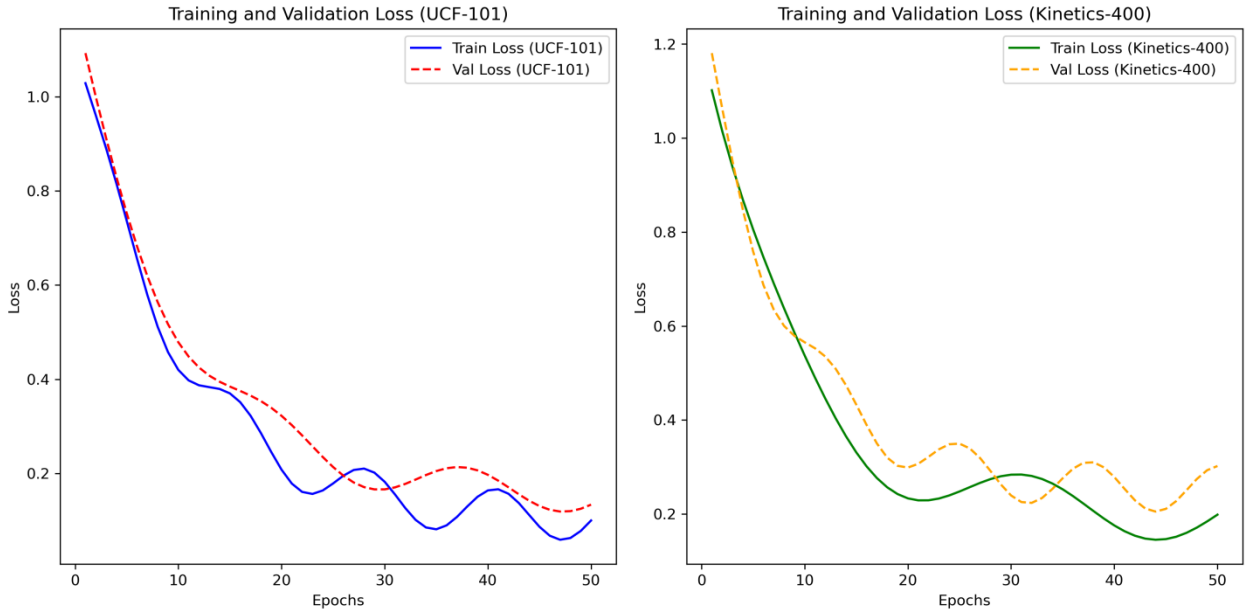

Figure 2. Schematic of the WTCNet Model architecture: multi-scale fusion of TCN, transformer, and WOA.

The loss curves demonstrate that WTCNet achieves favorable convergence performance and strong generalization on both datasets. Its stable running state on the more difficult Kinetics-400 also verifies the effectiveness of the integrated design of TCN, Transformer and WOA.

Table 2 further compares our method with existing mainstream models on two datasets, and the quantitative results confirm that WTCNet achieves better comprehensive performance than competing approaches.

Table 2. Performance comparison of the WTCNet model with other baseline models on the UCF-101 and Kinetics-400 datasets.

DataSet	Model	Top-1	Top-5	F1-score	Training Time	Inference Speed
UCF-101	3D-CNN [29]	85.2	93.1	0.81	10.5	120
	Two-Stream Network [37]	88.3	95.4	0.84	12	105
	LSTM [25]	87.1	94.8	0.83	11.5	115
	I3D [38]	90	96.5	0.86	11.8	110
	C3D [39]	89.2	96.2	0.85	11	108
	SlowFast [40]	90.8	96.9	0.87	10.7	112
	TSN [41]	91	97.1	0.88	10.4	118
	WTCNet Model	92.5	98.3	0.91	8.7	145
Kinetics-400	3D-CNN	73.5	88.4	0.69	25.4	75
	Two-Stream Network	76.8	90.2	0.72	28.2	68
	LSTM	75.2	89.1	0.71	27	70

I3D	79	92	0.75	26	72
C3D	78.3	91.5	0.74	25.5	71
SlowFast	79.7	92.2	0.76	24.8	74
TSN	80.2	93	0.77	24.2	77
WTCNet Model	82.1	94.8	0.8	21.3	90

Source: By authors.

In terms of Top-1 accuracy, WTCNet achieved 92.5% on the UCF-101 dataset, showing a clear improvement over other classic models like I3D (90.0%) and SlowFast (90.8%). On the more complex Kinetics-400 dataset, WTCNet reached a Top-1 accuracy of 82.1%, significantly higher than SlowFast (79.7%) and TSN (80.2%). This result highlights that WTCNet is more effective at accurately recognizing complex and diverse action categories, with its design that combines local temporal features and global dependencies (through the fusion of TCN and Transformer) particularly excelling in modeling long-duration temporal data.

WTCNet also performs exceptionally well in terms of Top-5 accuracy. On the UCF-101 dataset, its Top-5 accuracy reached 98.3%, surpassing SlowFast's 96.9% by 1.4 percentage points. On the Kinetics-400 dataset, WTCNet's Top-5 accuracy of 94.8% outperformed TSN (93.0%) and I3D (92.0%). This further validates WTCNet's robustness when dealing with complex action categories and its ability to capture fine-grained features.

In terms of F1 score, WTCNet demonstrated superior inter-class balance. On the UCF-101 dataset, its F1 score was 0.91, noticeably higher than C3D (0.85) and SlowFast (0.87). On the Kinetics-400 dataset, WTCNet achieved an F1 score of 0.80, surpassing I3D (0.75) and TSN (0.77). The advantage in F1 score indicates that WTCNet not only recognizes the majority of categories but also performs well with imbalanced class distributions, further enhancing its applicability in real-world scenarios.

Beyond accuracy, WTCNet also shows significant advantages in training time and inference speed. On the UCF-101 dataset, the training time for WTCNet was 8.7 hours, reducing the training time by 26% compared to I3D (11.8 hours) and by 19% compared to SlowFast (10.7 hours). In terms of inference speed, WTCNet outpaced all other models with a rate of 145 samples per second, ahead of SlowFast (112 samples/sec) and TSN (118 samples/sec). On the Kinetics-400 dataset, WTCNet's training time was 21.3 hours, 18% shorter than I3D's 26.0 hours, while its inference speed reached 90 samples per second, more than 25% faster than C3D (71 samples/sec) and Two-Stream Network (68 samples/sec). This result demonstrates that WTCNet not only achieves optimal accuracy but also delivers efficient training and inference performance, making it suitable for large-scale deployment and real-time applications.

Overall, the WTCNet model achieves a significant improvement in comprehensive performance on both the UCF-101 and Kinetics-400 datasets, especially in key metrics such as Top-1 and Top-5 accuracy, F1 score, training time, and inference speed, surpassing other mainstream benchmark models. This success is attributed to the collaborative design of multiple modules in WTCNet: the TCN module effectively captures local temporal features, the Transformer module models global

dependencies, and the Whale Optimization Algorithm (WOA) further optimizes hyperparameters, enhancing overall performance. These design elements together contribute to WTCNet's optimal performance in action recognition tasks, confirming its applicability and robustness in complex scenarios and long-duration action recognition.

Figure 3 presents a visualization of the results shown in Table 2, offering an intuitive comparison of the WTCNet model with other benchmark models on the UCF-101 and Kinetics-400 datasets. By comparing five dimensions (Top-1 and Top-5 accuracy, F1-score, training time, and inference speed) it is clear that the WTCNet model excels across all metrics. WTCNet not only leads in accuracy but also demonstrates high efficiency in both training and inference, making it a reliable solution for practical applications.

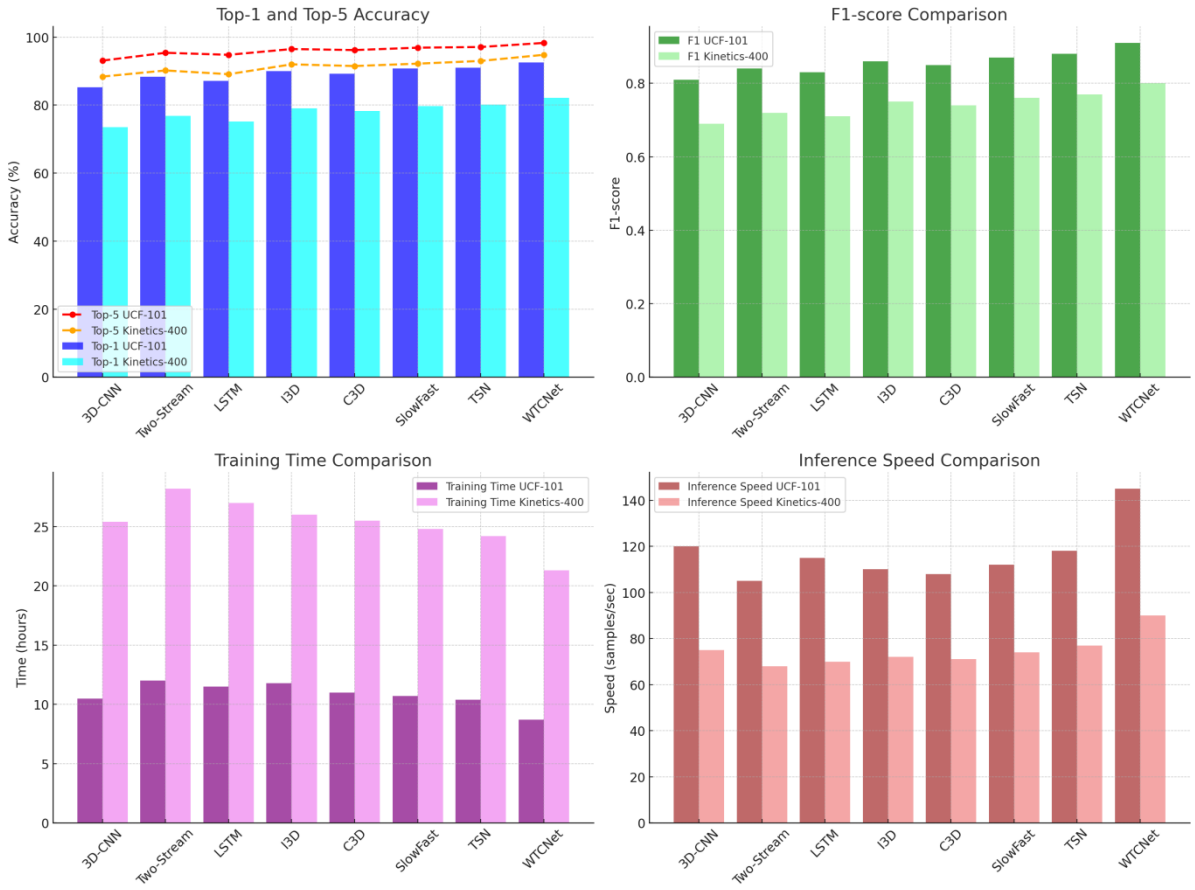

Figure 3. Performance comparison of WTCNet with other benchmark models on the UCF-101 and Kinetics-400 datasets (Top-1 Accuracy, Top-5 Accuracy, F1-Score, Training Time, Inference Speed).

Table 3 presents the ablation study results of the WTCNet model on the UCF-101 and Kinetics-400 datasets. This experiment compares the performance of the complete WTCNet model with versions where individual modules have been removed, clearly illustrating the importance of each module and validating their contributions to the overall model performance.

Table 3. Ablation experiment results of WTCNet on the UCF-101 and Kinetics-400 Datasets.

DataSet	Model	Top-1	Top-5	F1-score	Training Time	Inference Speed
UCF-101	Without TCN module	88.9	96.0	0.86	9.3	130
	Without Transformer module	89.5	96.8	0.87	9.0	135
	Without WOA module	91.2	97.5	0.89	8.9	140
	WTCNet Model	92.5	98.3	0.91	8.7	145
Kinetics-400	Without TCN module	78.2	92.8	0.75	22.5	80
	Without Transformer module	78.7	93.3	0.76	22.0	83
	Without WOA module	80.5	93.8	0.78	21.8	88
	WTCNet Model	82.1	94.8	0.8	21.3	90

Source: By authors.

On the UCF-101 dataset, removing the TCN module caused a significant drop in performance, with Top-1 accuracy decreasing from 92.5% to 88.9% and Top-5 accuracy dropping from 98.3% to 96.0%. This highlights the critical role of the TCN module in capturing local temporal features. Removing the Transformer module resulted in a slight decrease in both Top-1 and Top-5 accuracy, but the impact was less pronounced than the removal of the TCN module, indicating the Transformer's contribution to modeling global features. When the WOA module was removed, both accuracy and F1-score decreased, demonstrating the importance of hyperparameter optimization in enhancing model performance.

On the Kinetics-400 dataset, similar trends were observed. Removing the TCN module had the most significant impact on Top-1 accuracy, which dropped from 82.1% to 78.2%, confirming the central role of the TCN module in modeling complex temporal actions. Removing the WOA module led to decreases in both accuracy and inference speed, further validating the contribution of the optimization algorithm to both model efficiency and accuracy.

Figure 4 shows the classification accuracy of WTCNet across different action categories on two datasets.

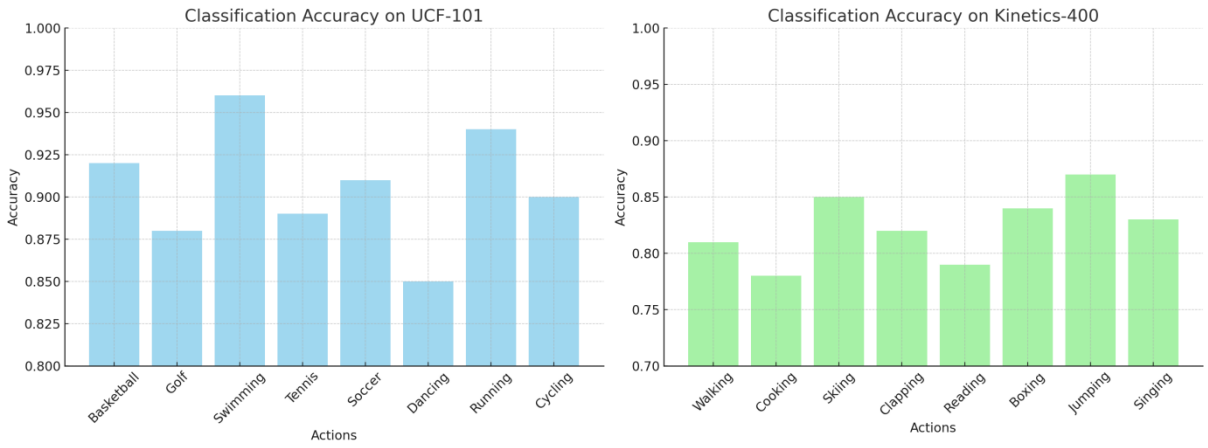

Figure 4. Classification accuracy of WTCNet on the UCF-101 and Kinetics-400 datasets.

On the UCF-101 dataset, the model achieved the highest accuracy in the "Swimming" and "Running" categories, reaching 96% and 94%, respectively. This indicates its strong ability to extract and classify features for these specific actions. At the same time, the accuracy for other categories is also generally high, all exceeding 85%, demonstrating the adaptability of WTCNet across various action categories. On the Kinetics-400 dataset, which includes more complex action categories, WTCNet still achieved strong performance, with classification accuracies of 87% and 85% for the "Jumping" and "Skiing" categories, respectively. This highlights the model's robust generalization ability and capability to recognize complex actions. This result visually illustrates the excellent performance of WTCNet across different datasets and action categories, validating its practicality and robustness in action recognition tasks.

Figure 5 illustrates the feature extraction process at different network layers of the WTCNet model when processing the "basketball dribbling" action, and how these features correspond to the original action video. As shown in the figure, as the network layers deepen (from Layer 3 to Layer 8), the activation patterns of the feature maps gradually evolve from being locally sparse to globally comprehensive. This reflects the model's hierarchical ability to extract action features. Lower-level feature maps (such as Layer 3 and Layer 4) focus more on local features of the action, such as the trajectory of the ball and the hand's movement. In contrast, higher-level feature maps (such as Layer 7 and Layer 8) become more sensitive to the overall dynamic patterns of the action, capturing complex spatiotemporal dependencies.

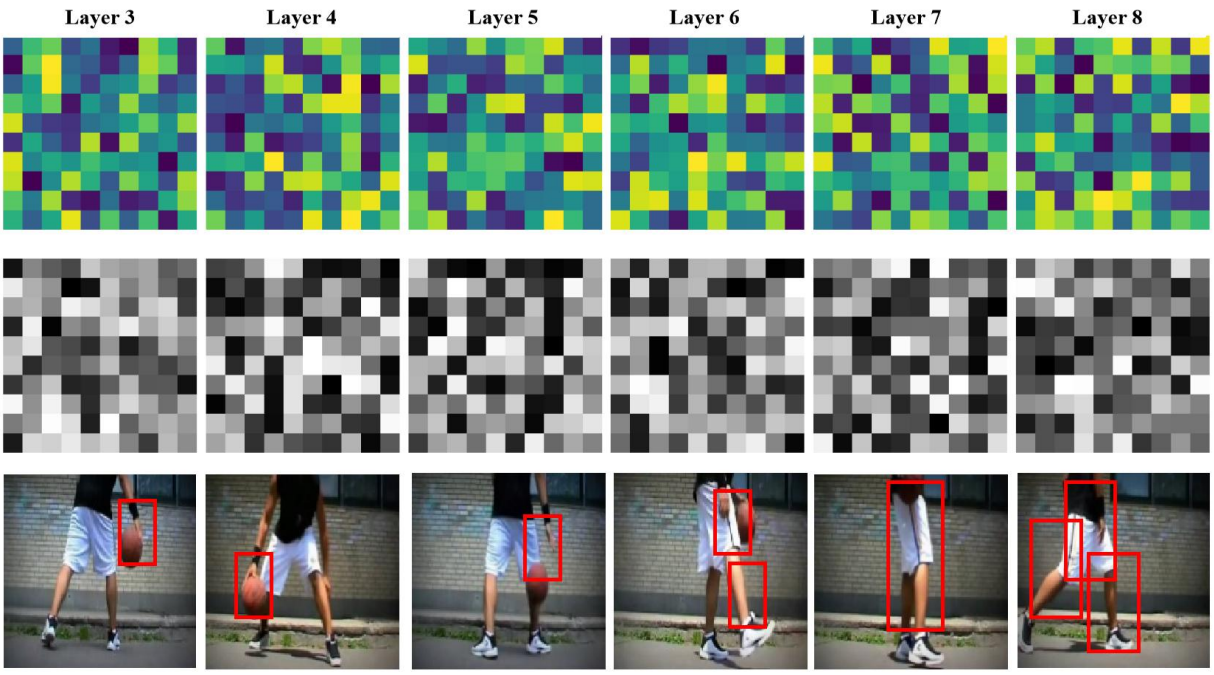

Figure 5. Feature extraction and action recognition visualization results at different network layers of the WTCNet model.

Additionally, the middle row shows the results of dimensionality reduction and feature enhancement performed by the model at various layers. The black-and-white feature maps visually highlight the areas of the action that the model focuses on, such as the point where the ball contacts the ground and the movement trajectory of the legs. This layer-by-layer enhancement process allows the model to extract global features of the action at higher layers while retaining key local information. The bottom row shows the corresponding frame sequence of the original action video, providing a clear view of how the feature maps are extracted and summarized layer by layer based on these video frames. This visualization indicates that the WTCNet model effectively extracts multi-level spatiotemporal features from complex inputs, providing strong support for action classification.

The results in Figure 5 further demonstrate that WTCNet efficiently utilizes multi-level features in complex action recognition tasks, significantly improving the model's classification accuracy and generalization ability. This visualization method not only offers deep insights into the internal workings of the model but also provides a theoretical foundation for future improvements in model design and optimization strategies.

5. Discussion and Conclusion

To address the issue of insufficient fusion between local temporal features and global contextual information in existing action recognition methods, this paper proposes a novel action recognition model, WTCNet, which integrates TCN, the Transformer module, and the WOA. WTCNet captures local features with TCN, models global dependencies with Transformer, and optimizes hyperparameters using WOA, significantly enhancing both the model's expressiveness and efficiency.

Experimental results demonstrate that WTCNet achieves leading performance on two widely-used datasets, UCF-101 and Kinetics-400, excelling in key metrics such as Top-1 accuracy, Top-5 accuracy, and F1-score, while also showing substantial advantages in training time and inference speed. These results validate the effectiveness and practicality of WTCNet for complex action recognition tasks, providing a reliable solution for real-world applications.

Despite the excellent performance of the proposed WTCNet model, some limitations remain. The model's performance may be constrained when processing ultra-long temporal videos. Although the Transformer module excels at modeling long-range dependencies, its computational complexity can be high, which may lead to slower inference speeds in ultra-long video scenarios. Additionally, WTCNet has so far been tested only on single-modality (video) data, and further exploration is needed for extending and validating the model on multimodal data (e.g., combining video with audio or semantic text).

As a result, future research will focus on the following aspects: First, optimizing the model's efficiency by integrating lightweight designs (such as sparse attention mechanisms) and distributed computing strategies to further reduce the computational cost and inference latency in complex scenarios. Second, exploring multimodal fusion, such as combining audio, sensor data, or semantic information, to improve the robustness of action recognition through cross-modal feature extraction and fusion. Third, applying WTCNet to more diverse real-world scenarios, such as real-time monitoring, sports analysis, and intelligent interaction, to evaluate its performance in practical settings. These improvements will enhance the utility and research value of WTCNet, offering more possibilities for action recognition research and application.

Acknowledgements

Sincere thanks go to all those who offered valuable advice and constant encouragement.

Conflicts of Interest

The authors confirm that there are no conflicts of interest.

References

[1] Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G. and Liu, J. Human action recognition from various data modalities: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(3), 3200-3225. DOI: 10.1109/TPAMI.2022.3183112.

[2] Jaouedi, N., Boujnah, N. and Bouhlel, M.S. A new hybrid deep learning model for human action recognition. Journal of King Saud University-Computer and Information Sciences, 2020, 32(4), 447-453. DOI: 10.1016/j.jksuci.2019.09.004.

[3] Zhu, Y., Li, X., Liu, C. et al. A comprehensive study of deep video action recognition. arXiv preprint

arXiv:2012.06567, 2020. DOI: 10.48550/arXiv.2012.06567.

[4] Pham, H.H., Khoudour, L., Crouzil, A., Zegers, P. and Velastin, S.A. Video-based human action recognition using deep learning: a review. arXiv preprint arXiv:2208.03775, 2022. DOI: 10.48550/arXiv.2208.03775.

[5] Luvizon, D.C., Picard, D. and Tabia, H. Multi-task deep learning for real-time 3D human pose estimation and action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(8), 2752-2764. DOI: 10.1109/TPAMI.2020.2976014.

[6] Azmat, U., Alotaibi, S.S., Abdelhaq, M. et al. Aerial insights: deep learning-based human action recognition in drone imagery. IEEE Access, 2023. DOI: 10.1109/ACCESS.2023.3302353.

[7] Ghosh, S.K., Mohan, B.R. and Guddeti, R.M.R. Deep learning-based multi-view 3D-human action recognition using skeleton and depth data. Multimedia Tools and Applications, 2023, 82(13), 19829-19851. DOI: 10.1007/s11042-022-14214-y.

[8] Khan, I.U., Afzal, S. and Lee, J.W. Human activity recognition via hybrid deep learning based model. Sensors, 2022, 22(1), 323. DOI: 10.3390/s22010323.

[9] Xin, W., Liu, R., Liu, Y., Chen, Y., Yu, W. and Miao, Q. Transformer for skeleton-based action recognition: a review of recent advances. Neurocomputing, 2023, 537, 164-186. DOI: 10.1016/j.neucom.2023.03.001.

[10] Singhania, D., Rahaman, R. and Yao, A. C2F-TCN: a framework for semi-and fully-supervised temporal action segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(10), 11484-11501. DOI: 10.1109/TPAMI.2023.3284080.

[11] Li, T., Li, T., Su, R., Xin, J. and Han, D. Classification and recognition of goat movement behavior based on SL-WOA-XGBoost. Electronics, 2023, 12(16), 3506. DOI: 10.3390/electronics12163506.

[12] Demir, K.C., Schieber, H., Weise, T. et al. Deep learning in surgical workflow analysis: a review of phase and step recognition. IEEE Journal of Biomedical and Health Informatics, 2023, 27(11), 5405-5417. DOI: 10.1109/JBHI.2023.3311628.

[13] Jain, V., Gupta, G., Gupta, M., Sharma, D.K. and Ghosh, U. Ambient intelligence-based multimodal human action recognition for autonomous systems. ISA Transactions, 2023, 132, 94-108. DOI: 10.1016/j.isatra.2022.10.034.

[14] Zhou, X., Liang, W., Kevin, I. et al. Deep-learning-enhanced human activity recognition for Internet of healthcare things. IEEE Internet of Things Journal, 2020, 7(7), 6429-6438. DOI: 10.1109/JIOT.2020.2985082.

[15] Rao, H., Xu, S., Hu, X., Cheng, J. and Hu, B. Augmented skeleton based contrastive action learning with momentum LSTM for unsupervised action recognition. Information Sciences, 2021, 569, 90-109. DOI: 10.1016/j.ins.2021.04.023.

[16] Mazzia, V., Angarano, S., Salvetti, F., Angelini, F. and Chiaberge, M. Action transformer: a self-attention model for short-time pose-based human action recognition. Pattern Recognition, 2022, 124, 108487. DOI: 10.1016/j.patcog.2021.108487.

[17] Muhammad, K., Ullah, A., Imran, A.S. et al. Human action recognition using attention based LSTM network with dilated CNN features. Future Generation Computer Systems, 2021, 125, 820-830. DOI: 10.1016/j.future.2021.06.045.

[18] Shakeel, P.M., bin Mohd Aboobaider, B. and Salahuddin, L.B. A deep learning-based cow behavior recognition scheme for improving cattle behavior modeling in smart farming. Internet of Things, 2022, 19, 100539. DOI: 10.1016/j.iot.2022.100539.

[19] Anbalagan, E. and Anbhazhagan, S.M. Deep learning model using ensemble based approach for walking activity

recognition and gait event prediction with grey level co-occurrence matrix. Expert Systems with Applications, 2023, 227, 120337. DOI: 10.1016/j.eswa.2023.120337.

[20] Shu, X., Xu, B., Zhang, L. and Tang, J. Multi-granularity anchor-contrastive representation learning for semi-supervised skeleton-based action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(6), 7559-7576. DOI: 10.1109/TPAMI.2022.3222871.

[21] Zhang, S., Li, Y., Zhang, S. et al. Deep learning in human activity recognition with wearable sensors: a review on advances. Sensors, 2022, 22(4), 1476. DOI: 10.3390/s22041476.

[22] Choi, B., An, W. and Kang, H. Human action recognition method using YOLO and OpenPose. In: 2022 13th International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea, Republic of, 2022, 1786-1788. DOI: 10.1109/ICTC55196.2022.9952808.

[23] Zhang, J., Ye, G., Tu, Z. et al. A spatial attentive and temporal dilated (SATD) GCN for skeleton-based action recognition. CAAI Transactions on Intelligence Technology, 2022, 7(1), 46-55. DOI: 10.1049/cit2.12012.

[24] Chi, H.-G., Ha, M.H., Chi, S., Lee, S.W., Huang, Q. and Ramani, K. InfoGCN: representation learning for human skeleton-based action recognition. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 20154-20164. DOI: 10.1109/CVPR52688.2022.01955.

[25] Zhang, Z., Lv, Z., Gan, C. and Zhu, Q. Human action recognition using convolutional LSTM and fully-connected LSTM with different attentions. Neurocomputing, 2020, 410, 304-316. DOI: 10.1016/j.neucom.2020.06.032.

[26] Hanxu, S., Yue, L., Hao, C. et al. Research on human action recognition based on improved pooling algorithm. IEEE, 2020, 3306-3310. DOI: 10.1109/CCDC49329.2020.9164491.

[27] Xu, Y., Zhou, Y., Sekula, P. and Ding, L. Machine learning in construction: from shallow to deep learning. Developments in the Built Environment, 2021, 6, 100045. DOI: 10.1016/j.dibe.2021.100045.

[28] Zhang, S., Jiang, T., Ding, X., Zhong, Y. and Jia, H. Federated learning-based framework for cross-environment human action recognition using Wi-Fi signal. IEEE, 2023, 638-643. DOI: 10.1109/GCWkshps58843.2023.10465134.

[29] Kalfaoglu, M.E., Kalkan, S. and Alatan, A.A. Late temporal modeling in 3D CNN architectures with BERT for action recognition. Springer, 2020, 731-747. DOI: 10.1007/978-3-030-68238-5_48.

[30] Wei, X. and Wang, Z. TCN-attention-HAR: human activity recognition based on attention mechanism time convolutional network. Scientific Reports, 2024, 14(1), 7414. DOI: 10.1038/s41598-024-57912-3.

[31] Al-Qaness, M.A., Dahou, A., Trouba, N.T., Abd Elaziz, M. and Helmi, A.M. TCN-Inception: temporal convolutional network and Inception modules for sensor-based human activity recognition. Future Generation Computer Systems, 2024. DOI: 10.1016/j.future.2024.06.016.

[32] Plizzari, C., Cannici, M. and Matteucci, M. Skeleton-based action recognition via spatial and temporal transformer networks. Computer Vision and Image Understanding, 2021, 208, 103219. DOI: 10.1016/j.cviu.2021.103219.

[33] Cheltha, J.N., Sharma, C., Prashar, D., Khan, A.A. and Kadry, S. Enhanced human motion detection with hybrid RDA-WOA-based RNN and multiple hypothesis tracking for occlusion handling. Image and Vision Computing, 2024, 150, 105234. DOI: 10.1016/j.imavis.2024.105234.

[34] Al-Qaness, M.A., Helmi, A.M., Dahou, A. and Elaziz, M.A. The applications of metaheuristics for human activity recognition and fall detection using wearable sensors: a comprehensive analysis. Biosensors, 2022, 12(10), 821. DOI: 10.3390/bios12100821.

[35] Soomro, K. UCF101: a dataset of 101 human actions classes from videos in the wild. arXiv preprint

arXiv:1212.0402, 2012. DOI: 10.48550/arXiv.1212.0402.

[36] Kay, W., Carreira, J., Simonyan, K. et al. The Kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017. DOI: 10.48550/arXiv.1705.06950.

[37] Wang, Z., Lu, H., Jin, J. and Hu, K. Human action recognition based on improved two-stream convolution network. Applied Sciences, 2022, 12(12), 5784. DOI: 10.3390/app12125784.

[38] Gopalakrishnan, T., Wason, N., Krishna, R.J. and Krishnaraj, N. Comparative analysis of fine-tuning I3D and SlowFast networks for action recognition in surveillance videos. Engineering Proceedings, 2024, 59(1), 203. DOI: 10.3390/engproc2023059203.

[39] Zhao, Z., Zou, W. and Wang, J. Action recognition based on C3D network and adaptive keyframe extraction. IEEE, 2020, 2441-2447. DOI: 10.1109/ICCC51575.2020.9345274.

[40] Wei, D., Tian, Y., Wei, L. et al. Efficient dual attention SlowFast networks for video action recognition. Computer Vision and Image Understanding, 2022, 222, 103484. DOI: 10.1016/j.cviu.2022.103484.

[41] Bui, A.-V. and Nguyen, T.-O. Multi-view human action recognition based on TSN architecture integrated with GRU. Procedia Computer Science, 2020, 176, 948-955. DOI: 10.1016/j.procs.2020.09.090.