Infrared Action Recognition Model Based on Improved ST-GCN

Xiaoliang Zhu, Ziwei Zhou

***Abstract*—Infrared imaging technology is capable of capturing the thermal radiation emitted by the human body in conditions with insufficient visible light. Consequently, infrared behavior recognition leverages this capability to detect and analyze human movements in low-light or complex environments. However, infrared images are often affected by noise interference, which can obscure target features. To tackle these challenges, we propose an infrared human behavior recognition model. Within this model, human regions in infrared images are detected by YOLOv8 and passed on to AlphaPose to predict the locations of skeletal keypoints in the human body. Subsequently, the acquired skeletal sequences are employed to predict actions in ST-GCN. Simultaneously, we introduced the LKA attention mechanism and the PReLU activation function for structural optimization within the ST-GCN. These improvements enabled the ST-GCN to extract action features from skeletal keypoints more effectively, thereby enhancing the accuracy of infrared behavior recognition.Through extensive ablation studies, we have demonstrated that our proposed LPST-GCN model significantly enhances the performance of infrared action recognition and achieves excellent results on both the UNISV dataset (99.02%) and the NTU RGB+D dataset (95.86%).**

***Index Terms*—Action recognition, infrared, alphapose, yolov8n, st-gcn**

I. Introduction

With the advancement of Industry 5.0, intelligent factory production lines are increasingly becoming the cornerstone of the manufacturing sector due to their high levels of automation and intelligence. However, in these advanced production environments, human workers continue to play an irreplaceable role in critical aspects such as assembly, quality control, and maintenance. Therefore, safeguarding the safety and health of these workers is particularly important.Behavior recognition technology serves as a vital tool for real-time monitoring and behavior detection, playing a crucial role in ensuring safety within intelligent production environments [1].

Although visible light devices are commonly employed in traditional surveillance systems, their performance can be significantly hindered in specialized factory settings. These devices rely on the reflection and refraction of light; thus, they may experience substantial degradation in image quality or even loss of information under harsh conditions such as inadequate lighting, smoke, dust, or elevated temperatures. Such challenges render it extremely difficult to utilize visible light equipment for precise measurement and observation within complex and dynamic industrial contexts.

In contrast, infrared cameras present an effective solution for addressing these challenges due to their distinct advantages. Infrared technology can detect thermal radiation emitted by objects, providing highly sensitive and clear images when visible light is inadequate or restricted. This capability renders infrared cameras invaluable for identifying equipment malfunctions, monitoring temperature anomalies, and ensuring overall safety. In specialized production environments, the use of infrared cameras not only exemplifies technological advancement but also serves as a crucial element in enhancing both production safety and efficiency.

Currently, research on action recognition predominantly concentrates on environments illuminated by visible light, while the domain of action recognition in infrared settings remains relatively underexplored. In recent years, some researchers have begun to investigate the application of infrared cameras within the field of action recognition. For instance, Gao et al. [2] put forward a method for recognizing human movements in infrared images by employing Convolutional Neural Networks (CNN). Wu et al. [3] introduced NIRExpNet, a three-stream 3D convolutional neural network model designed to address the challenges associated with Facial Expression Recognition (FER) under active near-infrared (NIR) illumination. Liu et al. [4], focusing on infrared human action recognition, underscored the importance of global temporal information in characterizing body part motion within videos. Chen et al. [5] proposed a method for temporal action detection utilizing infrared video. They generated the optical flow of infrared data by constructing a Flow Estimation Network (FEN) and optimized it in conjunction with the entire network architecture. Mehta et al. [6] developed a motion- and region-aware adversarial learning-based model for detecting fall events from infrared videos. Quan et al. [7] introduced a knowledge distillation method termed ARCTIC for RGB to infrared video action recognition, providing an effective approach to enhance the performance of infrared action recognition using RGB data.

Although the aforementioned methods have yielded some results in infrared action recognition, they primarily focus on end-to-end video classification tasks and still fall short in accurately recognizing individual actions within specific frames of a video. To address these limitations, we propose an action recognition algorithm based on infrared imaging

Manuscript received December 1, 2024; revised March 30, 2025.

Xiaoliang Zhu is a Postgraduate Student of School of Electronic Information, University of Science and Technology Liaoning, Anshan, 114051, P. R. China. (e-mail: 1097339974@qq.com).

Ziwei Zhou is an Associate Professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China.(Corresponding author, phone: 86-13941255680; e-mail: 381431970@qq.com).

for the effective identification of human behaviors in infrared environments. Given that the accuracy of top-down keypoint detection models is contingent upon the precision of target detection frames, we employ the YOLOv8n model for target detection to enhance the accuracy of keypoint detection related to human skeletons. Furthermore, we introduce an attention mechanism and a novel activation function into the spatio-temporal graph convolutional network to improve the model's recognition accuracy.

II. Infrared Human Movement Recognition Algorithm

In this study, we propose an infrared target-based action recognition algorithm comprising three primary components. Pedestrians in the infrared video are detected using the YOLOv8n model [8], which facilitates the extraction of human target regions. The human body region identified by YOLOv8n is subsequently input into the AlphaPose model to predict the locations of skeletal key points on the human body. The resulting skeletal sequences are then utilized for modeling within Spatio-Temporal Graph Convolutional Networks (ST-GCN) [10] to extract spatio-temporal features and accurately predict movements. However, challenges arise due to low resolution, high noise levels, and unclear features inherent in thermal infrared images, which hinder accurate extraction of human skeletal sequences in ST-GCN [11]. To address these limitations, this study introduces a novel node attention model termed LPST-GCN that integrates ST-GCN with Large Kernel Attention (LKA) [12] and employs the PreLU activation function [13]. LKA effectively captures long-range dependencies through a single convolution by utilizing a large kernel size (e.g., 7x7 or larger). This enhancement enables improved comprehension and modeling of temporal properties associated with actions, particularly those involving longer durations. Furthermore, by substituting the original ReLU activation function [14] with a parameterized PreLU activation function — known for its superior learning capabilities and enhanced feature extraction — the recognition accuracy of our model is significantly elevated. The processing flow of the proposed method is illustrated in Fig. 1.

A. Human Target Detection Based on Yolov8n

YOLOv8n, Ultralytics' target detection model released in the early 2023, is a model of lighter weight and has faster processing speeds [15]. This model builds upon the strengths of its predecessors while incorporating several enhancements in terms of model architecture, loss functions, and data augmentation techniques. These improvements have resulted in a substantial increase in performance across various tasks, including target detection, semantic segmentation, and image classification. The structure of the YOLOv8n network is illustrated in Fig. 1(a).

The YOLOv8n model introduces several innovative methods aimed at enhancing performance. The C2f (CSPLayer_2Conv) module serves as a foundational component of the backbone network, providing greater computational efficiency and reduced parameter redundancy compared to the C3 module utilized in YOLOv5. This C2f module not only maintains robust feature extraction capabilities by integrating depth-separable convolution with inflated convolution but also significantly decreases both the model's parameter count and computational cost. Such structural optimization renders YOLOv8 more suitable for real-time applications or scenarios characterized by limited resources, all while preserving high accuracy.Furthermore, YOLOv8n introduces the concept of a decoupled detection head, which decomposes complex detection tasks into multiple independent subtasks, each equipped with a specialized detection head. This divide-and-conquer strategy enhances the model's detection performance across targets of varying scales while effectively reducing its computational complexity.In terms of feature extraction, YOLOv8n improves upon the feature pyramid network (FPN), thereby bolstering the model's ability to detect human targets at different scales through a multi-scale feature fusion technique. The top-down pathway and lateral connections within the FPN structure adeptly integrate high-level semantic information with low-level spatial information, enriching hierarchical representations of features and enhancing their expressive capability.To further improve inference speed and memory efficiency, YOLOv8n employs model pruning and quantization techniques during the inference stage. These techniques effectively reduce both the number of model parameters and computational costs without significantly compromising accuracy, thus meeting stringent requirements for real-time performance and resource efficiency in practical applications.

B. Alphapose-based Human Pose Estimation

Currently, there are two main techniques for human pose estimation: top-down estimation and bottom-up estimation. In the top-down approach, the first step is to detect human positions from a series of images. Subsequently, pose estimation is carried out, and skeletal key points are extracted for each individual target. This method prioritizes the identification of individuals before analyzing their poses. On the other hand, the bottom-up estimation method simultaneously detects skeletal key points of all persons in video frames. These detected key points are then matched to construct a graph. Through a graph optimization process, any incorrect connections in the graph are eliminated to ensure the accuracy of pose estimation. Although the bottom-up method usually has a faster processing speed, it does not fully utilize the inherent global spatial information of human poses. Therefore, compared with the top-down method, its recognition accuracy is lower [16].

The two primary challenges faced by top-down methods in multi-person pose estimation are inaccurate bounding box localization and pose redundancy. To address these issues, AlphaPose introduces the RMPE framework, which consists of three main components: the Symmetric Spatial Transformer Network (SSTN), the Parametric Pose Non-Maximum Suppression (P-Pose NMS), and the Pose Guidance Proposal Generator (PGPG). Each component is specifically designed to tackle problems related to bounding

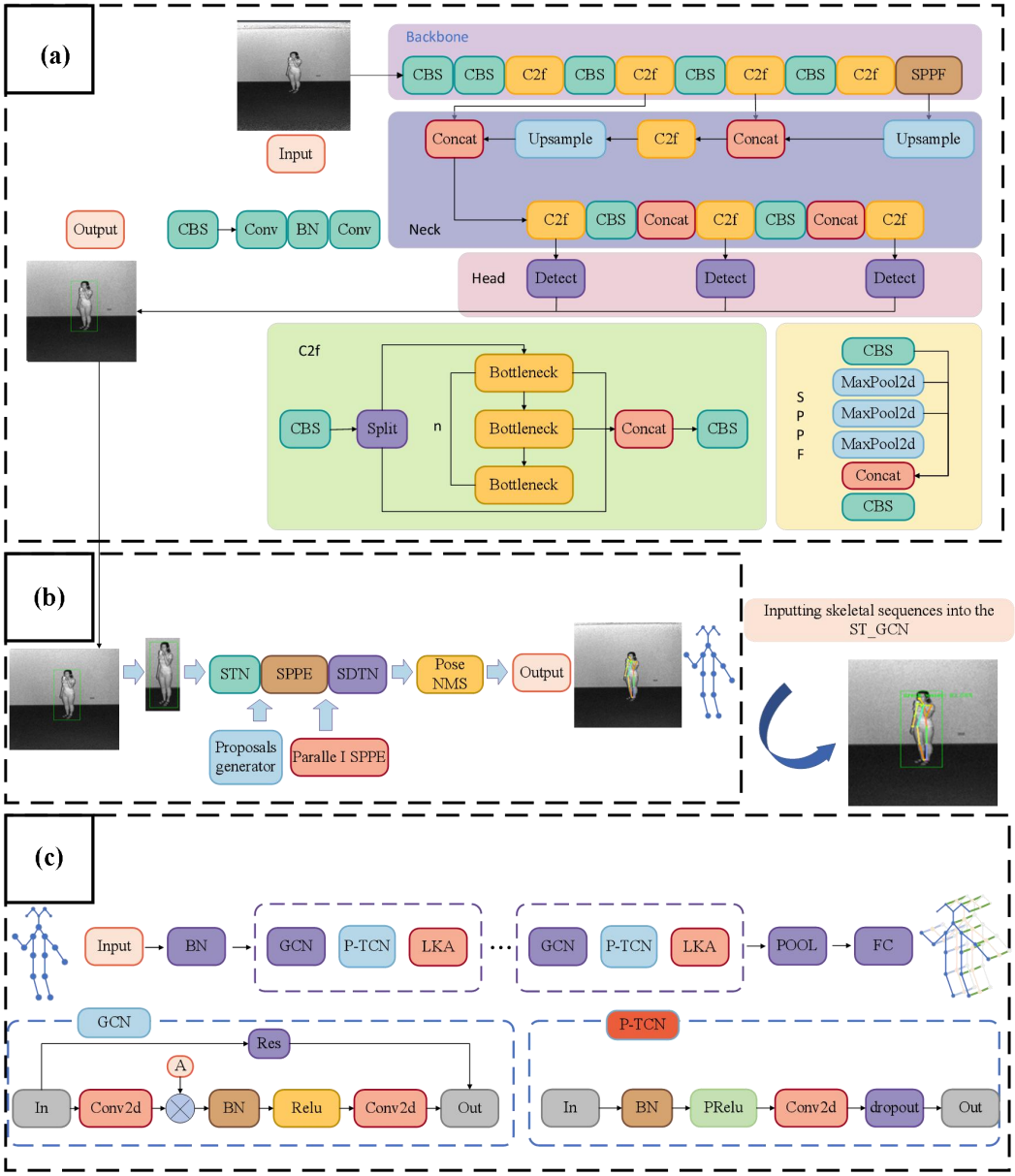

Fig. 1. General structure of the infrared behavior recognition model

box localization errors, pose redundancy, and training data augmentation. Fig. 1(b) illustrates the overall architecture and workflow of the AlphaPose framework.

The SSTN is developed to mitigate bounding box localization errors. It comprises a spatial transform network (STN), a stacked hourglass model for single-person pose estimation (SPPE), and an inverse spatial transform network (SDTN). The STN processes inaccurate input frames to obtain precise target candidate regions; the SPPE estimates human body poses; finally, the SDTN maps these estimated poses back to their original image coordinates. Additionally, parallel processing within SPPE enhances STN performance by returning larger error margins during training, thereby facilitating more accurate extraction of human detection frames.

P-Pose NMS aims to eliminate redundant poses while enhancing accuracy in human pose estimation. This method incorporates two criteria for elimination: confidence-based elimination and distance-based elimination. When either criterion is satisfied, redundant poses are removed effectively. This approach addresses redundancies that arise from independent operations of multiple bounding boxes.

The PGPG component is employed to augment training samples by synthesizing a substantial amount of training data through simulated offsets in prediction frames. This strategy enables improved effectiveness in training for the SSTN+SPPE module, ultimately enhancing the accuracy of prediction frames generated by the target detector.

C. Improving the St-gcn Network

The backbone of the Spatio-Temporal Graph Convolutional Network (ST-GCN) consists of 10 ST-GCN units. The first four layers are designed with 64 output channels, the subsequent three layers contain 128 output channels, and the final three layers feature 256 output channels. As depicted in Figure 2, each ST-GCN unit comprises a spatial convolutional layer, a temporal convolutional layer, and a residual structure , and the original TCN structure. Within the spatio-temporal graph convolutional unit, a learnable edge weight parameter is utilized to evaluate the significance of edges connecting nodes. In the temporal convolutional layer of the ST-GCN, a single kernel-sized convolution operation is performed using a fixed architecture. Successive joint information serves as input features for the ST-GCN network; these

input features undergo batch normalization before being processed through a series of ST-GCN cell layers for convolution operations and are ultimately pooled using global average pooling. The input features are represented as a four-dimensional matrix (N, C, T, V), where N denotes the number of videos, C represents the number of joint features, T indicates the number of key frames, and V signifies the number of joints. The improved ST-GCN is presented in Fig. 1(c).

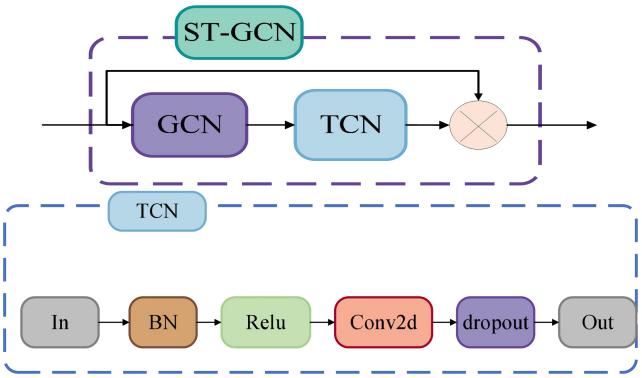

Fig. 2. Original ST-GCN Cell Layer and Original TCN Structure Diagrams

Since the temporal features of actions often span extended periods, and ST-GCN primarily learns local features within a specific neighborhood without adequately capturing relevant information across all nodes (global information), its capacity to capture long-distance dependencies is inherently limited. This limitation adversely impacts the model's recognition accuracy. To address this issue, this paper proposes a novel node attention model that integrates ST-GCN units with Large Kernel Attention (LKA), an attention mechanism specifically designed to preserve both channel and spatial dimensions. This approach enhances the significance of cross-latitude interactions while diminishing the influence of less important features. When implemented as a modified LSTGCN model, the model has the ability to capture longer distance dependencies within a single convolution through the use of a large kernel convolution (7x7). This enhancement enables the model to better comprehend and represent the temporal properties of actions, particularly for those with extended durations. Furthermore, it successfully combines global and local features to construct rich hierarchical structures.

The Large Kernel Attention (LKA) module is a mechanism that effectively integrates spatial attention and channel attention. Its primary function is to generate an attention map through large-kernel convolution, thereby emphasizing key target regions. The underlying formula can be expressed as follows:

$$\text{Attention Map} = LargeKernelConv(x) \quad (1)$$

$$\text{Output} = \text{Attention Map} \otimes \text{Input Feature} \quad (2)$$

The Input Feature represents the input feature map, while $\otimes$ denotes the elemental product. However, large kernel convolutions are often associated with high computational costs and an increased number of parameters. To address this issue, the LKA module ingeniously decomposes the large kernel convolution into three components, aiming to alleviate the computational burden and reduce the number of parameters: (1) Depth Separable Convolution (DW-Conv), which is utilized to capture local spatial features; (2) Depth-widening Convolution (DW-D-Conv), which focuses on capturing long-range dependencies; and (3) Channel Convolution (Conv1×1), which manages channel information. This decomposition process, illustrated in Fig. 6, effectively minimizes the computational cost of the LKA module while preserving its capability to generate attention mapping.

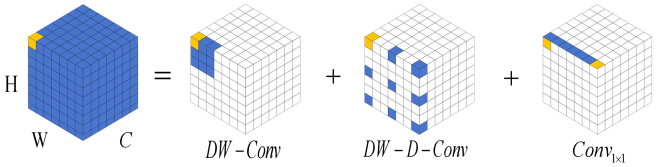

Fig. 3. Decomposition Diagram of Large Kernel Convolution

The Input Feature denotes the input feature map, while $\otimes$ represents the elemental product. However, large kernel convolutions are frequently associated with significant computational costs and an increased number of parameters. To mitigate this challenge, the LKA module adeptly decomposes the large kernel convolution into three distinct components, aiming to alleviate both the computational burden and parameter count: (1) Depth Separable Convolution (DW-Conv), which is employed to capture local spatial features; (2) Depth-widening Convolution (DW-D-Conv), which concentrates on capturing long-range dependencies; and (3) Channel Convolution (Conv1×1),

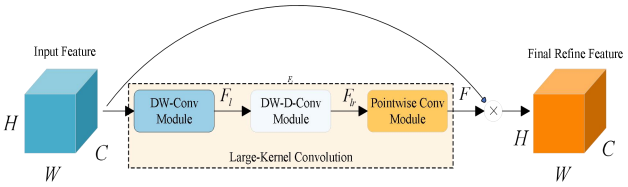

Fig. 4. Large Kernel Attention (LKA) Module

which processes channel information. This decomposition process, as illustrated in Fig. 3, effectively reduces the computational cost of the LKA module while maintaining its ability to generate attention mapping.

The specific formula is derived as illustrated in Fig. 4, under the assumption that the initial input feature map is represented as $x \in R^{C\times H\times W}$, where c, h and w denote the number of channels, height, and width, respectively. Initially, the feature map x undergoes processing through a depthwise convolution (DW-Conv) submodule that employs a 5 × 5 convolutional kernel with a stride of 2. This operation results in a local attention feature map denoted as $F_l \in R^{C\times H\times W}$ (refer to Equations 3).

Firstly, x traverses the DW-Conv submodule, which encompasses a 5×5 convolution kernel, a 2×2 stride convolution, and outputs a local attention feature map, designated as $F_l \in R^{C\times H\times W}$ (Equation 3).

Subsequently, the local attention feature map F_l is passed through a depthwise separable convolution (DW-D-Conv) submodule utilizing a 7×7 convolution kernel with a stride of 9 and an expansion factor of 3. This step produces another

feature map that captures long-distance dependencies, which we denote as $F_{lr} \in R^{C\times H\times W}$ (see Equation 4). Following this process, the long-range dependency feature map F_{lr} is subjected to further processing via a 1×1 convolution to generate the channel-adaptive attention map represented by $F \in R^{C\times H\times W}$ (Equation 5). Finally, this attention map F is multiplied by the initial feature map x, resulting in the refined feature representation expressed as $Attention = F\otimes$x (Equation 6). This procedure effectively enhances key features while suppressing non-key features, thereby improving both the model's expressive capability and its generalization performance.

$$F_l \in R^{C\times H\times W} \tag{3}$$

$$F_{lr} \in R^{C\times H\times W} \tag{4}$$

$$F \in R^{C\times H\times W} \tag{5}$$

$$Attention = F\otimes \text{x} \tag{6}$$

To enhance the accuracy of the network and address the issue of gradient vanishing, which may arise when the negative domain of the Rectified Linear Unit (ReLU) is zero during extensive training, this paper employs the Parametric Rectified Linear Unit (PReLU) to optimize the activation function of Temporal Convolutional Networks (TCN) within the Spatio-Temporal Graph Convolutional Network (ST-GCN), as illustrated in Fig. 8. During training, PReLU has the capability to learn the slope parameter associated with ReLU, thereby improving model performance. The definition of PReLU is as follows:

$$f(y_i) = \begin{cases} y_i & y_i > 0 \\ a_i y_i & y_i \le 0 \end{cases} = \max(0, y_i) + a_i \min(0, y_i) \tag{7}$$

Where y_i represents the input on the i-th channel, and a_i denotes the coefficient that governs the slope of the negative portion of the activation function. It is crucial to note that the subscript i in a_i indicates that the slope of the negative segment of the PReLU activation function can vary across different channels, thereby providing enhanced flexibility. The parameter a_i of PReLU can be trained in the same manner as that of other layers via backpropagation and an optimizer. furthermore, their updating formulas can be derived using the chain rule. Assuming that at a particular layer, the gradient of a_i is:

$$\frac{\partial E}{\partial a_i} = \sum_{y_i} \frac{\partial E}{\partial f(y_i)} \frac{\partial f(y_i)}{\partial a_i} \tag{8}$$

In the expression provided as (2), the objective function is denoted by the symbol ε, while its partial derivative $\partial E/f(y_i)$ represents the gradient that is propagated through the deep network. The gradient of the activation function can be derived from equation (3).

$$\frac{\partial f(y_i)}{\partial a_i} = \begin{cases} 0 & y_i > 0 \\ y_i & y_i \le 0 \end{cases} \tag{9}$$

Specifically, the gradient can be expressed as follows:

$$\frac{\partial E}{\partial a} = \sum_i \sum_{y_i} \frac{\partial E}{\partial f(y_i)} \frac{\partial f(y_i)}{\partial a_i} \tag{10}$$

When the input value exceeds 0, the gradient is equal to 0; conversely, when the input value is less than or equal to 0, the gradient is determined by a specific formula. In this study, we employ a channel-sharing strategy, whereby all channels within each network layer share identical parameters. This approach significantly reduces the number of variables that need to be introduced. In this context, the gradient of a parameter can be defined as the summation of the gradients from all channels in a given layer. Consequently, we are able to update and optimize the parameters more efficiently.

III. Experimental Results and Analysis

A. Dataset

All data utilized in the YOLOv8n experiments were obtained from Google Datasets. In this study, a dataset designated as Dataset 1 was constructed for infrared image data.The enhanced ST-GCN network employed the UNISV infrared dataset [17] as Dataset 2 for training and validation purposes. Each video was segmented into individual frames, and the key joints of the human body within each frame were extracted using the AlphaPose algorithm. Subsequently, each frame is annotated in accordance with the action category and stored in pkl file format as the training data for the enhanced ST-GCN network.A total of ten distinct actions were selected: walk, squat, singlewave, shakehands, pushpeople, jump,jogging,fight,embrace,doublewave.

Dataset-1 is fabricated through the labeling of manually screened and data-enhanced images.The techniques employed for data augmentation included the following: 1) mirroring of images; 2) random rotation of angles; 3) random adjustments to image contrast; and 4) the addition of Gaussian noise. Ultimately, a total of 5,893 images were acquired. Image annotation was performed using LabelImg software to label the collected data, which encompasses information regarding location, category, recognition difficulty, and additional attributes. All data will be converted into PASCAL VOC format.

Dataset-2 is composed of the UNISV dataset.As illustrated in Figure 5, the UNISV dataset provides a representative example of the infrared surveillance video content. The UNISV dataset was constructed using night-time infrared surveillance videos and encompasses ten different human behavior categories. During the dataset's construction phase, original or minimally edited video samples were selected to ensure authenticity and diversity. In the compilation process, factors such as sample diversity and environmental complexity were fully considered. The recorded behaviors cover various scenarios and were all captured in outdoor environments, which reflect typical installation locations of surveillance cameras in the real world. All recordings were conducted at night to fully utilize the infrared function. Fifteen individual participants with different anthropometric characteristics, including height and body type, were involved in generating the dataset. These participants performed predefined actions in various scenarios, ensuring that the dataset contains a wide range of motion manifestations. This approach helps evaluate action

recognition algorithms under realistic surveillance conditions while maintaining the controllability of experimental parameters.

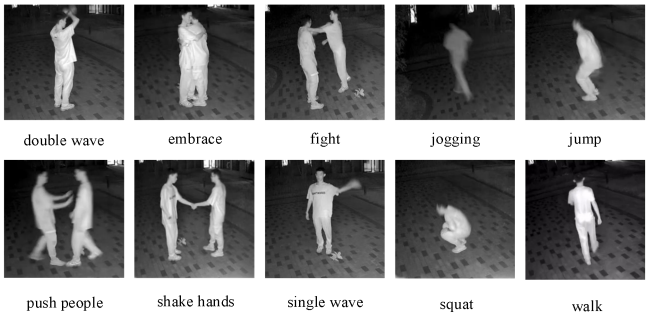

Fig. 5. Samples of action from UNISV dataset

B. Experimental Platform and Model Evaluation

The model training platform operates on a Windows system that is equipped with an NVIDIA GeForce RTX 4060 graphics card and employs the PyTorch deep learning framework. The model testing platform consists of laptops featuring Intel® Core™ i7-14700HX CPUs.

Target detection experiments were conducted utilizing a custom-built infrared dataset, designated as Dataset-1. A variety of YOLO algorithms were employed for comparative analysis on this dataset, with each algorithm being trained over 200 epochs. The evaluation of the algorithms was based on four criteria to determine which one achieves the optimal balance between detection accuracy and processing speed. The results of these experiments are presented in the accompanying Table I.

From the experimental results, it is clear that YOLOv8n demonstrates superior detection accuracy and enhanced detection speed. In comparing the lightweight versions of each YOLO model, YOLOv8n stands out with the fastest detection speed and the fewest model parameters. Therefore, we have selected YOLOv8n as our thermal imaging human detection model due to its high accuracy, rapid detection capabilities, and lightweight design.

TABLE I
COMPARISON WITH OTHER MODELS ON DATASET-1

Model	Params/ M	mAP/%	GFLOPs/G	ModelSize/ MB
YOLOv3-tiny	10.12	66.7	13.5	17.4
YOLOv5n	7.21	68.3	15.8	14.4
YOLOv7-tiny	6.03	64.2	13.1	12.3
YOLO8n	3.09	74.5	8.1	6.3

Target behavior recognition experiments were conducted utilizing the thermal infrared video dataset, referred to as Dataset-2. During the training phase, the number of epochs was set to 100, with Stochastic Gradient Descent (SGD) employed as the optimizer. The learning rate was established at 0.01, momentum was set to 0.9, and weight decay was configured at 1e-5.

To identify the optimal integration point for incorporating the LKA attention mechanism, a series of experiments were performed in this study. These experiments were categorized into ten groups; Group 1 represented the original STGCN network, while the remaining nine groups reflected results from experiments where LKA attention was integrated into various GCN+TCN modular layers of the STGCN architecture. The findings from these ablation experiments are summarized in Table II. As indicated in Table II, integrating LKA attention within Layer 5 resulted in superior performance, achieving an accuracy rate that is 1.56% higher than that of the original STGCN model. Therefore, in this paper, we will incorporate LKA attention in Layer 5 to construct LSTGCN.

TABLE II
COMPARISON OF DIFFERENT ST-GCN LAYERS FOR ADDING LKA ON THE DATASET-2

Group number	Model	Accuracy%	Loss%
1	St-GCN	0.9564	0.1063
2	St-GCN1+Lka	0.9607	0.1033
3	St-GCN2+Lka	0.9428	0.1111
4	St-GCN3+Lka	0.9528	0.1085
5	St-GCN4+Lka	0.9695	0.1009
6	St-GCN5+Lka	0.9720	0.1014
7	St-GCN6+Lka	0.9516	0.1113
8	St-GCN7+Lka	0.9657	0.1021
9	St-GCN8+Lka	0.9491	0.1095
10	St-GCN9+Lka	0.9610	0.1062

Our final model represents an optimization of the proposed LTGCN framework, wherein the activation function has been replaced with the Parametric Rectified Linear Unit (PReLU) function. The original Spatio-Temporal Graph Convolutional Network (ST-GCN) employs the Rectified Linear Unit (ReLU) activation function, which effectively addresses the vanishing gradient problem associated with s-curves. However, the output of ReLU to negative inputs is zero, which can easily lead to neuronal “death”.To mitigate this limitation, we have substituted ReLU with PReLU, which provides enhanced learning capabilities and more effective feature representation. PReLU allows for the gradient values of the negative half-axis to be transformed into dynamically learnable parameters, thereby offering greater flexibility in adjusting these parameters throughout the training process to achieve optimal results. The experimental results presented in the Table III indicate that our model utilizing PReLU achieves an accuracy improvement of 1.82% compared to the original model employing ReLU.

TABLE III
RESULTS OF ABLATION EXPERIMENTS ON THE DATASET-2

Group number	Model	Accuracy%	Loss%
1	St-GCN	0.9564	0.1063
2	St-GCN5+Lka	0.9720	0.1014
3	St-GCN5+Lka+Prule	0.9902	0.0950

Figure 6 illustrates the changes in experimental accuracy over time. Overall, the Ours model demonstrates a faster convergence rate compared to the ST-GCN model,

indicating that the enhancements made to the original architecture facilitate more effective learning. The curve representing the ST-GCN model exhibits pronounced local fluctuations during the early stages of training, suggesting instability in its performance. In contrast, as training progresses into later stages, both models' curves tend to stabilize; however, the average AUC value for the Ours model remains slightly higher and displays a smoother trajectory. This observation underscores the advantages of our proposed model throughout the entire training process. The experimental results indicate that in terms of accuracy, the LPST-GCN model has achieved an improvement of 3.38 relative to the original STGCN model.

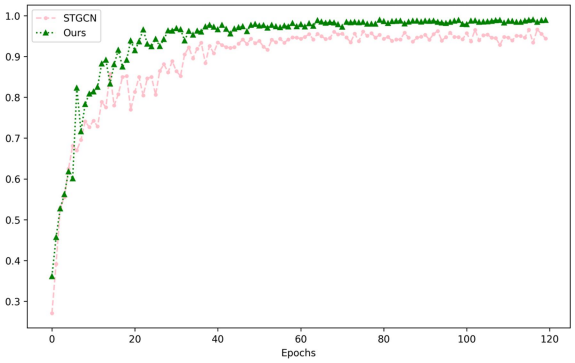

Fig. 6. Accuracy on the Dataset-2

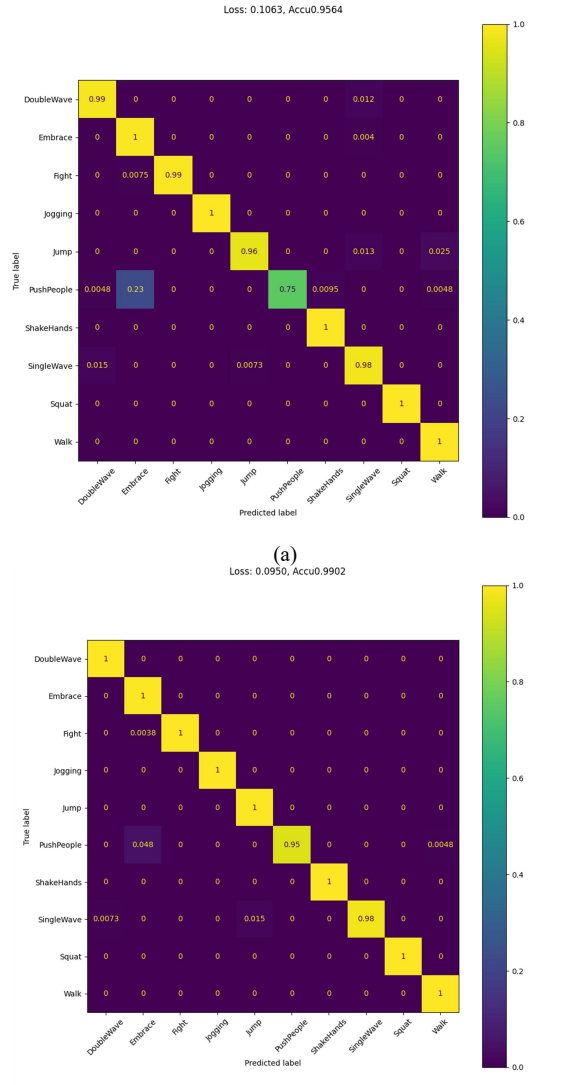

Fig. 7. Confusion Matrix

Fig. 7(a) depicts the original model, and Fig. 7(b) presents the modified one, with the confusion matrix employed to identify the ten operational categories with maximum accuracy.

Our model exhibits superior recognition accuracy, reflecting its predictive performance across each action category. Most action categories are predicted correctly with high probability, indicating that the model excels in this task. Although there is some degree of confusion among a few categories, it does not significantly affect the overall accuracy. The results suggest that the enhanced model achieves a high recognition rate and demonstrates strong robustness.

TABLE IV
COMPARISON WITH MAINSTREAM MODELS

Group number	Model	Accuracy%
1	I3D[18]	76.70%
2	SlowFast[19]	84.20%
3	MDJ+TPN[17]	91.40%
4	YOLOv3-Alphapose-ST-GCN	94.05%
5	Ours	99.02%

To further validate the efficacy of our proposed model, we compared the research method employed in this study with the more typical behavior recognition methods in recent years, and the results are presented in Table Ⅳ. The proposed modeling method achieves a recognition accuracy that is 7.62% higher than the MDN+TPN [17] method, 14.82% higher than the SlowFast [19] network, and 22.32% higher than the I3D model [18]. These results demonstrate that our algorithm significantly outperforms existing methods in detecting human actions in infrared environments.

TABLE V
VERSATILITY AND ROBUSTNESS VERIFICATION EXPERIMENTS ON NTU RGB+D DATASET

Group number	Model	Accuracy%
1	YOLOv3-Alphapose-ST-GCN	90.15%
2	Ours	95.86%

To further verify the universality and robustness of the LPST-GCN algorithm, the baseline algorithm LPST-GCN was tested on the NTU RGB+D public dataset. NTU RGB+D is a comprehensive action recognition dataset developed by Nanyang Technological University (NTU) in Singapore. It comprises 56,880 video samples across 60 distinct action categories. These actions were performed by 40 participants from three different viewpoints, ensuring both diversity and breadth in the dataset. Each sample includes recordings from three modalities: RGB video, depth map, and infrared image, providing rich multimodal resources for research. For dataset partitioning, two approaches are employed: Cross-Subject (CS) and Cross-View (CV). The CS approach divides the dataset into two groups based on participant IDs, facilitating the separation of training and test sets. Conversely, the CV approach partitions the dataset according to camera views to ensure balanced data distribution from various perspectives during training and testing. Given the large scale and high

hardware requirements of the NTU RGB+D dataset[20], this study focuses on validating methods using a subset of the infrared video clips corresponding to specific action categories. The results are shown in Table V.

IV. Conclusion

An enhanced ST-GCN model, in conjunction with the AlphaPose pose estimation algorithm, is proposed for intelligent monitoring and recognition of behaviors under poor lighting conditions. This approach aligns with the advanced concept of smart environments and contributes to reducing major safety incidents. The main contributions and experimental findings of this study are as follows: Existing research has primarily focused on action recognition in well-lit environments, often neglecting the complexities involved in accurately analyzing human motion under poor lighting conditions.A top-down, high-precision AlphaPose pose estimation model is employed to detect key points of the human skeleton within image sequences.An improved YOLOv8 model is utilized for human target detection, addressing off-target detection issues and enhancing the model's robustness against interference in complex environments.We propose a fusion of a spatio-temporal graph-based convolutional neural network (ST-GCN) with an LKA attention module, which not only captures local features but also enhances the understanding of global information.The ReLU activation function is replaced with the PReLU activation function to improve learning capabilities and feature representation.After optimization, our training network was evaluated on a self-constructed dataset, achieving an accuracy of 99.02%, representing a 3.38% improvement over the original model.

References

[1] W. J. Weijin, Y. S. Yongxia, H. Z. Haoran, P. C. Pingping, W. Z. Wanqing, and J. C. Junpeng, "Surveillance video behavior recognition mechanism based on ST-GCN under edge-cloud collaborative computing," *J. Nanjing Univ. (Nat. Sci.)*, vol. 58, pp. 163-174, 2022.

[2] C. Gao, Y. Du, J. Liu, J. Lv, L. Yang, D. Meng, and A. G. Hauptmann, "Infar dataset: Infrared action recognition at different times," *Neurocomputing*, vol. 212, pp. 36-47, 2016.

[3] Z. Wu, T. Chen, Y. Chen, Z. Zhang, and G. Liu, "NIRExpNet: Three-stream 3D convolutional neural network for near infrared facial expression recognition," *Applied Sciences*, vol. 7, no. 11, p. 1184, 2017.

[4] Y. Liu, Z. Lu, J. Li, T. Yang, and C. Yao, "Global temporal representation based cnns for infrared action recognition," *IEEE Signal Processing Letters*, vol. 25, no. 6, pp. 848-852, 2018.

[5] X. Chen, C. Gao, C. Li, Y. Yang, and D. Meng, "Infrared action detection in the dark via cross-stream attention mechanism," *IEEE Transactions on Multimedia*, vol. 24, pp. 288-300, 2021.

[6] V. Mehta, A. Dhall, S. Pal, et al., "Motion and region aware adversarial learning for fall detection with thermal imaging," *in Proceedings of the 25th International Conference on Pattern Recognition*, 2021, pp. 6321-6328.

[7] Z. Quan, Z. Zhenzhen, et al., "ARCTIC: A knowledge distillation approach via attention-based relation matching and activation region constraint for RGB-to-Infrared videos action recognition," *Comput. Vis. Image Underst.*, vol. 237, p. 103853, 2023.

[8] R. Varghese and M. Sambath, "YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness," *in Proc. 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, April 2024, pp. 1-6.

[9] H. -S. Fang et al., "AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time," *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7157-7173, 1 June 2022

[10] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," *in Proc. AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, April 2018.

[11] T. Si, F. He, P. Li, and X. Gao, "Tri-modality consistency optimization with heterogeneous augmented images for visible-infrared person re-identification," *Neurocomputing*, vol. 523, pp. 170-181, 2023.

[12] M. H. Guo, C. Z. Lu, Z. N. Liu, M. M. Cheng, and S. M. Hu, "Visual attention network," *Computational Visual Media*, vol. 9, no. 4, pp. 733-752, 2023.

[13] S. Zhang, J. Lu, and H. Zhao, "Deep network approximation: Beyond ReLU to diverse activation functions," *Journal of Machine Learning Research*, vol. 25, no. 35, pp. 1-39, 2024.

[14] R. C. Pinto and A. R. Tavares, "PReLU: Yet Another Single-Layer Solution to the XOR Problem," *arXiv preprint arXiv*:2409.10821, 2024.

[15] X. Li and Y. Zhang, "A Lightweight Method for Road Damage Detection Based on Improved YOLOv8n," *Engineering Letters*, vol. 33, no. 1, pp. 114-123, 2025.

[16] H. Li, R. Yuan, J. Chen, Q. Li, and C. Hu, "Research on Double Attention Mechanism High-resolution Network for Human Pose Estimation," *Engineering Letters*, vol. 33, no. 2, pp. 338-347, 2025.

[17] Z. Feng, X. Wang, J. Zhou, et al., "MDJ: A Multi-Scale Difference Joint Keyframe Extraction Algorithm for Infrared Surveillance Video Action Recognition," *Digital Signal Processing*, vol. 148, p. 104469, 2024.

[18] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," *in Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299-6308.

[19] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast Networks for Video Recognition," *in Proc. IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6202-6211.

[20] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," *in Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1010-1019.

XIAOLIANG ZHU was born in Tianjin Province, P. R. China, obtained his bachelor's degree in electrical engineering from Liaoning University of Science and Technology in 2022.
He is currently pursuing the M.S. degree in Electronic Information with University of Science and Technology Liaoning, Anshan, P. R. China. His research interest pertain to machine vision.

ZIWEI ZHOU was born in Liaoning Province, P. R. China, Associate Professor, Master's Degree Supervisor, received his Bachelor's and Master's Degrees from Liaoning University of Science and Technology in 1997 and 2007, and his Ph.D. Degree from Harbin Institute of Technology in 2013, with a major research interest in Artificial Intelligence, 3D Vision, Deep Learning, and Robotics Systems Research.