Machine Vision and Applications (2024) 35:65
https://doi.org/10.1007/s00138-024-01550-2

RESEARCH

Thermal infrared action recognition with two-stream shift Graph Convolutional Network

Jishi Liu[1] · Huanyu Wang[2] · Junnian Wang[1] · Dalin He[1] · Ruihan Xu[1] · Xiongfeng Tang[1]

Received: 11 November 2023 / Revised: 4 February 2024 / Accepted: 22 April 2024

Abstract
The extensive deployment of camera-based IoT devices in our society is heightening the vulnerability of citizens' sensitive information and individual data privacy. In this context, thermal imaging techniques become essential for data desensitization, entailing the elimination of sensitive data to safeguard individual privacy. Meanwhile, thermal imaging techniques can also play a important role in industry by considering the industrial environment with low resolution, high noise and unclear objects' features. Moreover, existing works often process the entire video as a single entity, which results in suboptimal robustness by overlooking individual actions occurring at different times. In this paper, we propose a lightweight algorithm for action recognition in thermal infrared videos using human skeletons to address this. Our approach includes YOLOv7-tiny for target detection, Alphapose for pose estimation, dynamic skeleton modeling, and Graph Convolutional Networks (GCN) for spatial-temporal feature extraction in action prediction. To overcome detection and pose challenges, we created OQ35-human and OQ35-keypoint datasets for training. Besides, the proposed model enhances robustness by using visible spectrum data for GCN training. Furthermore, we introduce the two-stream shift Graph Convolutional Network to improve the action recognition accuracy. Our experimental results on the custom thermal infrared action dataset (InfAR-skeleton) demonstrate Top-1 accuracy of 88.06% and Top-5 accuracy of 98.28%. On the filtered kinetics-skeleton dataset, the algorithm achieves Top-1 accuracy of 55.26% and Top-5 accuracy of 83.98%. Thermal Infrared Action Recognition ensures the protection of individual privacy while meeting the requirements of action recognition.

Keywords Action recognition · Thermal infrared · YOLOv7-tiny · Alphapose · Spatial-temporal graph convolutional networks · Two-stream structure

1 Introduction

The Internet of Things (IoTs) is profoundly transforming our society, enabling communication from any location at any time. The increasing deployment of IoT devices is contributing significantly to the improved management of public infrastructures and the enhanced well-being of citizens. Action recognition, a vital component of computer vision, is playing a crucial role in this context, helping identify ongoing actions within video clips captured by a wide range of camera-based IoT devices. In general, action recognition holds a significant role across various domains, encompassing human-computer interaction, monitoring and analyzing [1].

However, the extensive utilization of camera-based IoT devices heightens the vulnerability of sensitive information of citizens and individual data privacy. Thermal imaging techniques, before behavior recognition, substitutes sensitive information within the data to maintain individual privacy. Therefore, there is an increasing focus on infrared-based methods for human action recognition.

On the other hand, camera-based Industrial Internet of Things (IIoT) devices are more and more frequently operated in challenging lighting environments. These devices, designed to monitor and analyze various industrial processes, often find themselves deployed in warehouses, factories, and production facilities where natural light is limited. By considering that IIoT has positioned itself as a highly promising technology for shaping the future of Industry 5.0, it is nec-

✉ Huanyu Wang
huanyu@hnust.edu.cn

1 School of Physics and Electronic Science, Hunan University of Science and Technology, Xiangtan 411199, Hunan, China

2 School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 411199, Hunan, China

Published online: 13 May 2024

essary to go one step to further investigate to which extent thermal imaging techniques can be used for action recognition in IIoT domain.

Existing research predominantly focuses on visible light environments for action recognition [2], overlooking the complexities of accurately analyzing human movements in poor lighting conditions. Thermal imaging technique is a viable solution to this problem due to its unique imaging characteristics in infrared video. It provides a certain level of perspective and performs well in low-light environments, including night, haze, rain, and snow. The current approaches used for infrared action recognition are two-stream CNN [3], two-stream 3D CNN [4], CNN based on global temporal representation [5], infrared and visible image matching with Transform adversarial network [6], infrared action recognition based on cross-stream attention mechanism [7], etc. The majority of these approaches primarily focus on end-to-end video classification tasks, often falling short in their ability to accurately recognize individual actions within specific frames of the video. Consequently, these methods may not fully meet the requirements of contemporary monitoring tasks that demand fine-grained action recognition capabilities.

In our study, we present a novel lightweight algorithm for thermal imaging-based action recognition. The proposed approach begins by extracting the human body region location and skeletal information using the well-known YOLOv7-tiny [8] and Alphapose [9] framework, respectively. Subsequently, the obtained skeletal sequence is employed for modeling within the spatial-temporal Graph Convolutional Networks (ST-GCN) [10], enabling the extraction of spatial-temporal features and accurate prediction of the action. While promising in visible light environments, this method has limited efficacy in thermal infrared environments [11]. Thermal infrared images pose challenges for accurate human skeletal sequence extraction in ST-GCN due to low resolution, high noise, and indistinct features [12]. The model's use of target detection and pose estimation models from visible light datasets exacerbates these issues, leading to suboptimal results in action recognition [13, 14]. We propose a method to train the model using self-made infrared datasets, to address the aforementioned issue of imprecise skeleton extraction in thermal infrared images. Next, we select images from multiple infrared datasets and videos, label them using Labelme [15] for target detection and COCO-Annotator [16] for pose estimation. In addition, the conventional structure of ST-GCN exhibits several drawbacks. Firstly, it is computationally intensive. Secondly, the spatial and temporal graph sense fields are predetermined using heuristics, which limits their flexibility. Lastly, it lacks the capture of motion information in consecutive frames. We propose an innovative approach known as the two-stream shift map convolutional network (2s-ShiftGCN) [17]. The 2s-ShiftGCN addresses these limitations by incorporating two distinct streams: a static stream that captures spatio-temporal features related to the appearance of frames [18] and a motion stream that focuses on capturing spatio-temporal features pertaining to inter-frame motion [19].

ST-GCN training is not constrained to infrared spectral data by leveraging the skeleton information from Alphapose, as skeleton information remains unaffected by spectral variations [20]. For joint training, we trained YOLOv7-tiny and Alphapose with the thermal infrared dataset, and ST-GCN with the visible light dataset. This joint training approach aimed to create a robust model capable of accurate action recognition in thermal imaging conditions. We produced a human detection dataset (OQ35-human) and a skeletal dataset (OQ35-keypoint) in the infrared spectrum with the idea of joint training. They were used to train YOLOv7-tiny and Alphapose. We evaluated the similarity between the extracted skeleton from the thermal infrared image obtained by the model trained with the two datasets and the truth skeleton. We overlaid a thermogram on the skeletal joints and computed the cosine similarity to quantify the comparison. The obtained similarity score is 0.949, indicating a strong resemblance between the predicted and actual skeletons.

Additionally, we compared ST-GCN and 2s-ShiftGCN using two separate datasets, assessing their respective advantages and drawbacks. They are the custom thermal infrared action dataset (InfAR-skeleton) and the filtered kinetics-skeleton dataset [21]. The former of them is a thermal infrared dataset and the latter is a visible light dataset. The experimental results of 2s-ShiftGCN on the InfAR-skeleton dataset show higher Top-1 accuracy (88.06% vs. 76.59%) and Top-5 accuracy (98.28% vs. 96.03%) compared to ST-GCN. On the Filtered Dynamics Skeleton dataset, the algorithm achieves a Top-1 accuracy of 55.26% and a Top-5 accuracy of 83.98%, both higher than ST-GCN's Top-1 accuracy of 44.33% and Top-5 accuracy of 76.53%.

2 Related work

Action recognition falls into two primary categories: traditional techniques and deep learning approaches. Traditional approaches typically entail extracting high-dimensional visual features locally from video regions, which are then combined to form fixed-size video-level representations. These representations are subsequently fed into classifiers for final prediction [22]. Deep learning methods for action recognition can be classified into single-stream, two-stream, and skeleton-based approaches, based on the feature extraction techniques employed.

Karpathy et al. [23] proposed to use of 2D pre-trained convolution to fuse temporal information of consecutive frames to do prediction. LRCN [24] integrates convolu-

tional and LSTM layers to capture both spatial and temporal information for action recognition; C3D [25] employs 3D convolutional layers to process video data and extract spatio-temporal features; MiCT [26] combines 2D and 3D convolution modules; OFF [27] can extract spatial-temporal information simultaneously, especially the temporal information between frames.

The two-stream method introduces the fusion of spatial and temporal streams, with the spatial network capturing the spatial dependencies within the video and the temporal network capturing the temporal dynamics of the motion within the video [28]. TSN [29] has enhanced the two-stream architecture by adopting sparse clips instead of random sampling for video input. HiddenTwoStream [30] employs an unsupervised architecture to generate optical streams for all frames, enhancing the performance of the two-stream method. Instead of utilizing a single 3D network, I3D [21] employs distinct 3D networks in two separate streams. T3D [31] and a pre-trained 2D convolutional network undergo supervised migration learning, utilizing frames and clips from the same or different videos for knowledge transfer.

The skeleton-based approach lacks human appearance features compared to the traditional image input method. However, the skeleton sequence exhibits three distinct advantages. Firstly, the strong interdependence among individual nodes and their neighboring nodes enables the skeleton to capture rich information about the body structure. Secondly, temporal continuity not only presenting within the same joints, but also in the overall body structure. This contributes to the analysis of motion dynamics. Lastly, there is a mutually beneficial relationship between the spatial and temporal domains, allowing for comprehensive understanding of actions. These inherent advantages make skeleton-based action recognition a promising avenue for further advancement and development in the field.

ST-GCN proposed by Yan et al. [10] leverages a time-series representation of human joint positions to model the dynamic skeleton. It extends graph convolution to a spatial-temporal graph convolutional network for capturing spatial-temporal variations in joint relationships. The graph in this approach consists of two types of edges: spatial edges representing the natural connections between joints, and temporal edges connecting the same joints across consecutive time frames. Building upon this, a multi-layer spatial-temporal graph convolution is constructed to integrate information across both spatial and temporal dimensions. However, ST-GCN exhibits two drawbacks. Firstly, it is computationally intensive. Secondly, the spatial and temporal graph sense fields are predetermined using heuristics, which limits their flexibility.

Cheng et al. [17] proposed Shift-GCN to address the above drawbacks. The shift convolution operator serves as an inspiration for the proposed approach, which involves combining the 1×1 convolution operator with the Shift operation. This integration enables the 1×1 convolution to effectively fuse information from both spatial and channel domains to lead to an enhanced representation capturing capabilities. Shift-GCN applies non-local shift graph convolution and adaptive temporal shift graph convolution. The non-local shift graph convolution overcomes the constraint of physical intrinsic connectivity by transforming a single-frame skeletal graph into a fully connected graph. As a result, each node establishes a direct correlation with every other node in the graph. Adaptive temporal shift graph convolution method addresses the margin issue caused by integer realization by introducing a relaxation of the time shift parameter from an integer constraint to a real constraint through interpolation computation. This significantly enhances the generalization capability of the temporal shift graph convolution.

However, existing works have limited utilization of inter-frame motion information. This results in a significant loss of human motion information between adjacent frames and the lack of acquisition of motion representation information. To address this issue, this paper introduces the 2s-ShiftGCN method, which effectively captures both appearance and motion information from skeletal sequences in static as well as consecutive frames.

3 Thermal infrared human action recognition algorithm

The thermal infrared human action recognition algorithm in this paper comprises three primary components. Firstly, the YOLO algorithm is employed for human target detection in thermal imaging videos, assisted by a tracking algorithm. The target tracking algorithm employs a Kalman filter, a recursive algorithm utilized for estimating both target position and velocity. It is usually used for linear motion models of targets and has a certain tolerance for noise. Secondly, the detected human region locations are used for pose estimation to obtain the skeletal sequence. Lastly, GCN modeling skeletal sequences for action recognition.

3.1 YOLO-based human target detection

YOLOv7-tiny is the chosen model for human target detection due to its capabilities in meeting the requirements for accuracy, speed, and efficiency. The structure of the YOLOv7-tiny network is illustrated in Fig. 1.

YOLOv7-tiny adds the following methods to further improve the performance. Firstly, an efficient coupling network is employed to enhance the overall efficiency. YOLOv7 proposes an extended version of E-ELAN [32], based on ELAN [33]. This boosts the network's learning capacity

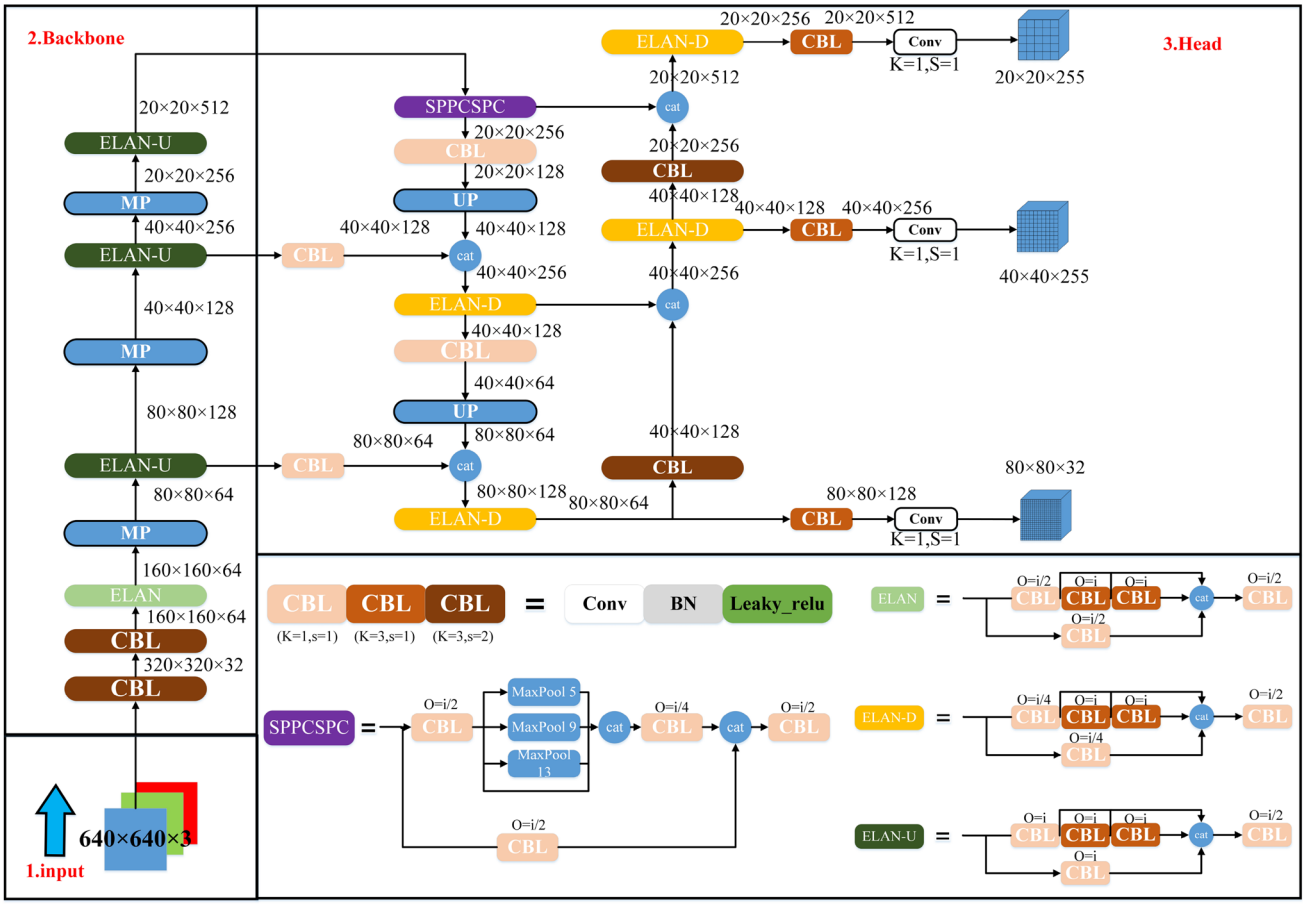


Fig. 1 YOLOv7-tiny network structure [8]

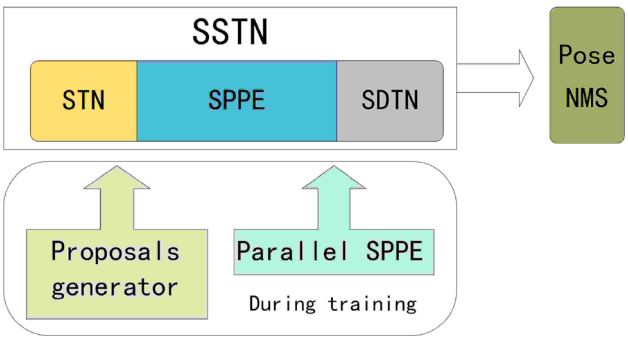


Fig. 2 RMPE framework

without impacting the original gradient pathways, regardless of their length or the number of computation modules. Additionally, the composite model is scaled to optimize its performance. Furthermore, convolutional reparameterization is introduced and improved to enhance the model's representational power. Lastly, an auxiliary training module is integrated to guide the label assignment strategy, allowing for a more refined and accurate prediction process [34].

3.2 Alphapose-based human posture estimation

Two primary challenges faced by top-down approaches in multi-person pose estimation are inaccurate bounding box localization and pose redundancy. Alphapose introduces the RMPE framework to address these issues. The framework consists of three components, each designed to address a specific problem [9]. The Symmetric Spatial Transformer Network (SSTN) handles bounding box positioning errors, the Parametric Pose Non-Maximum Suppression (P_Pose NMS) eliminates redundant poses, and the Pose-Guided Proposals Generator (PGPG) leverages intensive training data. Figure 2 depicts the overall architecture and workflow of the RMPE framework. SSTN consists of STN, SPPE and SDTN, where STN processes the inaccurate input boxes to get accurate target candidate regions, SPPE gets the estimated pose, and SDTN maps the estimated pose back to the original image coordinates. There is also a parallel SPPE, which serves to optimize the STN by returning a larger error for incorrect estimation.

NMS is employed to enhance the accuracy of human pose estimation by eliminating redundant poses. It operates by

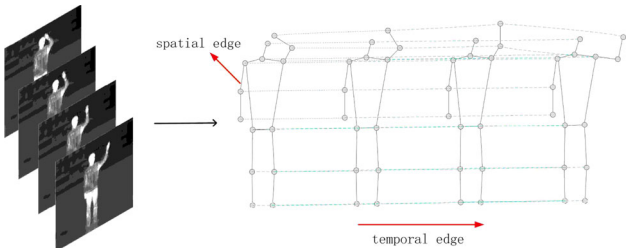


Fig. 3 Skeletal spatial-temporal graph

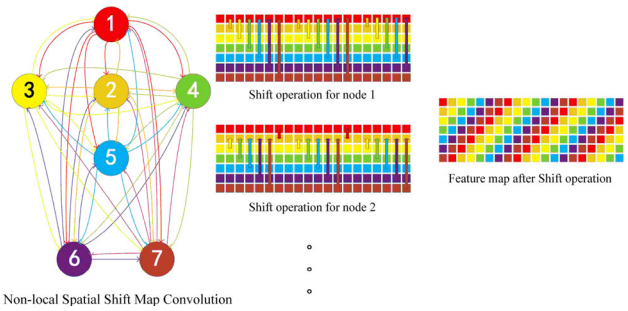


Fig. 4 Non-local shift graph convolution operation

searching for local maxima and suppressing non-maxima. Parametric non-maxima suppression involves two elimination criteria: confidence elimination and distance elimination. Redundant poses are eliminated if they meet either criterion.

PGPG performs data augmentation to train the SSTN + SPPE module, and the goal of Alphpose is to correct the prediction boxes generated by the target detector to overlap with the labeled boxes of each individual as much as possible through the SSTN + SPPE module. To effectively train this module, a substantial quantity of prediction boxes is necessary. These can be synthesized by modeling the offsets of the prediction boxes.

3.3 Action recognition based on Graph Convolutional Networks

As mentioned above, we construct a spatial-temporal graph $G = (V, E)$ to model a multi-layer skeletal sequence [10]. Figure 3 illustrates the skeletal spatial-temporal graph represented by a set of joint node matrices V. These matrices cover all human body joints across time and space, denoted as $V = V_{ti} | t = 1, 2, \ldots, T, i = 1, 2, \ldots, N$. Here, T represents the total number of video frames, and N represents the number of joints in each frame. The feature vector $F(v_{ti})$ is defined as $F(v_{ti}) = ((x, y), s)$, where (x, y) represents the coordinate values of the joint and s denotes the estimated confidence score. E encompasses both spatial and temporal edges. The spatial side collection E_s is defined as $E_s = v_{ti} v_{tj} | (i, j) \in H$, where H signifies the collection of target human joints in each frame. On the other hand, the temporal side collection E_t consists of the same node connections between different frames and represents the motion trajectory of each specific joint across frames.

ST-GCN divides the neighbor set of a node into three parts: the root node, the centripetal group, and the centrifugal group. The root node signifies the node itself, the centripetal group includes neighboring nodes closer to the skeleton's center of gravity than the root node in spatial position, while the centrifugal group contains neighboring nodes farther from the skeleton's center of gravity than the root node in spatial position [10]. The spatial convolution model in GCN is constructed using Eq. 1. The GCN network applies spatial convolution to the neighboring nodes of the skeleton graph, capturing local spatial information. Meanwhile, the TCN networks learn local temporal features by analyzing inter-frame joint changes.

$$X_{\text{out}} = \sum_{p \in P} M_{st}^{(p)} \circ \widetilde{A}^{(p)} X_{in} W_{st}^{(p)} \tag{1}$$

In the aforementioned equation, P represents one of the root nodes, centripetal group, or centrifugal group. X_{in} denotes the input data, $\widetilde{A}^{(\text{p})}$ is the normalized adjacency matrix, $M_{\text{st}}^{(\text{p})}$ corresponds to the network weights, and $W_{\text{st}}^{(\text{p})}$ represents the attention mechanism weights.

Two main drawbacks arise from dividing the points within the perceptual field into three subsets. Firstly, it leads to excessive computational complexity. Secondly, the predefined perceptual field of both temporal and spatial graphs imposes limitations on the extraction of action features. The traditional spatial-temporal graph convolution may restrict the ability to capture crucial action patterns. Despite ST-GCN extracting joint features from adjacent skeletal connections, it largely neglects distantly structured joints that may contain significant action patterns. Cheng et al. [17] introduced a novel approach, namely Shift-GCN, to address these aforementioned limitations. Shift-GCN replaces heavy called regular graph convolution with a novel shift-graph operation, providing flexible perceptual fields for spatial and temporal graphs. It also utilizes a lightweight point-by-point convolution to reduce computational complexity. Shift-GCN employs non-local shift graph convolution operation to eliminate inherent connectivity constraints and transform the skeletal graph into a fully connected graph. Each skeletal joint in the feature map $F \in \mathbb{R}^{N \times C}$ is directly connected to other skeletal joints based on the translation $d = i \ mod \ N$, resulting in a spiral-like feature structure, as illustrated in Fig. 4.

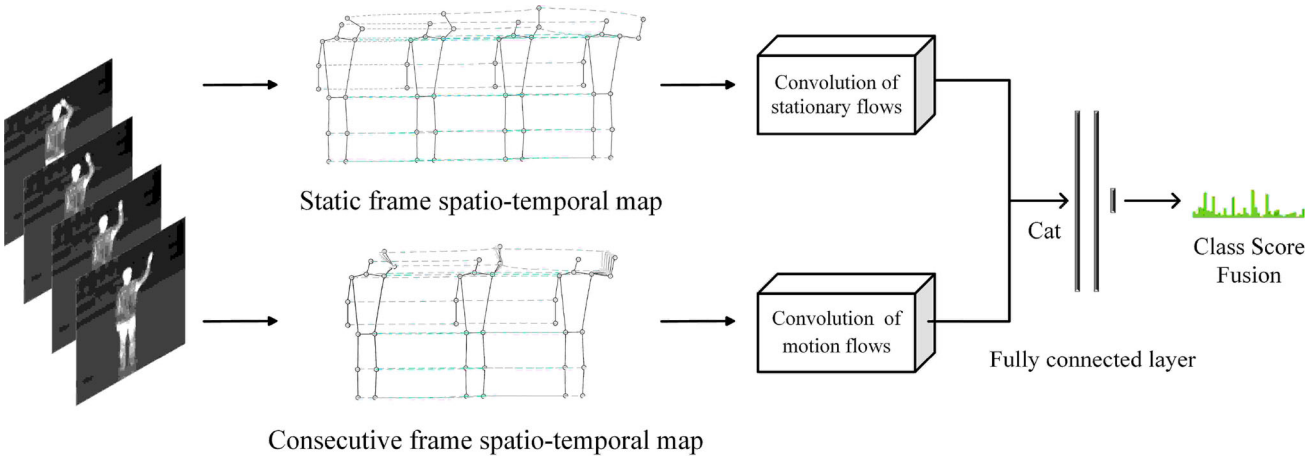


Fig. 5 2s-ShiftGCN structure

3.4 Two-stream shift Graph Convolutional Network

Shift GCN considers the stacked feature layers generated over time series, represented by the symbol $F \in \mathbb{R}^{T\times N\times C}$ for the spatiotemporal feature map with $F = \{F^1, F^2, \ldots, F^T\}$. This feature map can integrate traditional Shift convolution operators. When applying shift convolution to spatiotemporal feature maps, it is necessary to divide the channels evenly into 2u + 1 regions, each with offsets ranging from -u to u. If the offset area exceeds the channel length, the exceeding part will be discarded, and the vacated portion will be filled with zeros. Shift GCN proposes an adaptive temporal shift graph convolution without specifying the size of u. For each channel, a time offset parameter S_i needs to be learned. If the parameter is an integer, it can't transfer gradients. Therefore, it is necessary to relax the integer constraint and convert it to real numbers, followed by linear interpolation for calculation. The formula for linear interpolation is shown in Eq. 2.

$$\widetilde{F}_{(v,t,i)} = (1-\lambda)\cdot F_{(v,\lfloor t+S_i\rfloor,i)} + \lambda F_{(v,\lfloor t+S_i\rfloor+1,i)} \tag{2}$$

where $\lambda = S_i - \lfloor S_i\rfloor$ is the residual due to the realization of the integers, which needs to be compensated by means of interpolation. The anchor point is situated between the intervals $[\lfloor t+S_i\rfloor, \lfloor t+S_i\rfloor+1]$ after discretization. Therefore, interpolation is performed within this interval to estimate the desired values.

In this paper, we propose a two-stream shift Graph Convolutional Network (2s-ShiftGCN) to address the limitation of Shift-GCN in capturing spatial-temporal features of inter-frame motion [34]. The 2s-ShiftGCN comprises two streams, each focusing on a different representation. One stream captures the appearance of static frames, utilizing 3 channels to represent the skeleton information x, y, and s, where x and y denote the node coordinates and s denotes the confidence level. The other stream captures the motion between consecutive frames, utilizing 2 channels to represent the difference in node coordinates x, y between adjacent frames. The overall architecture of the s-ShiftGCN is shown in Fig. 5. The overall architecture of the 2s-ShiftGCN is shown in Fig. 5.

The motion history image is a natural outcome of capturing the joint motion between frames using optical flow techniques [35]. A compact representation of motion can be generated by utilizing the optical flow. This is achieved through frame-level connections to generate the motion flow frame. The motion spatial-temporal feature map is defined as the linear difference between consecutive frames, The subtraction of nodes between consecutive frames results in a decrease of 1 in the frame count, where the blank channels are filled with zeros. The motion flow input for consecutive frames can be expressed using Eq. 3. The feature extraction and fusion process is depicted in Fig. 6.

$$D_t((x,y),s) = F_t((x,y),s) - F_{t-1}((x,y),s) \tag{3}$$

The grayscale values of corresponding pixel points in the two frames are recorded as $F_t((x,y),s)$ and $F_{t-1}((x,y),s)$. Where (x,y) represents the joint's coordinates, and s denotes the estimated confidence score.

The input bone sequence has a shape of N×C×T×V×M, where N represents the batch size, C denotes the number of channels, T represents the frame count, V indicates the number of human nodes, and M signifies the number of input samples. After applying the 2s-ShiftGCN, we obtain two tensors with an output shape of $(N, 256)$. The two tensors are spliced and fused to get a tensor with an output shape of $(N, 256*2)$. The fusion algorithm concatenates multiple feature maps along the depth dimension to obtain a more enriched representation of features. Subsequently, the tensor is fed into a fully connected layer, resulting in a shape of (N, nc), where nc represents the number of action types. The network structure is depicted in Fig. 7.

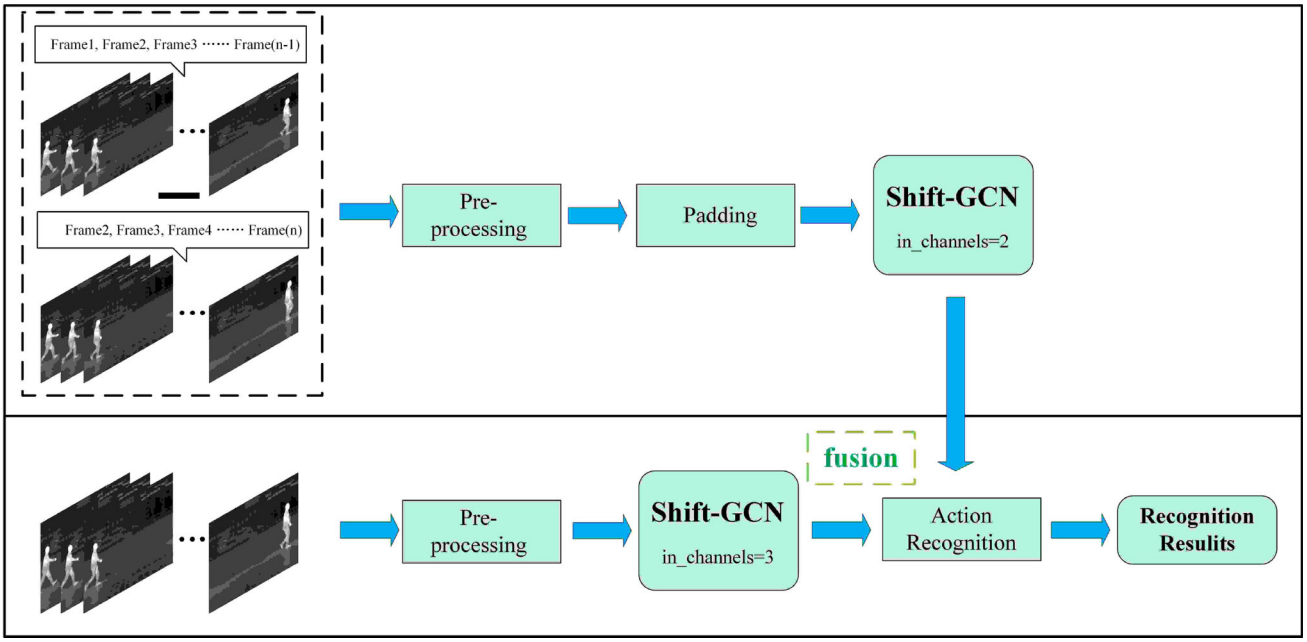


Fig. 6 Feature extraction and fusion

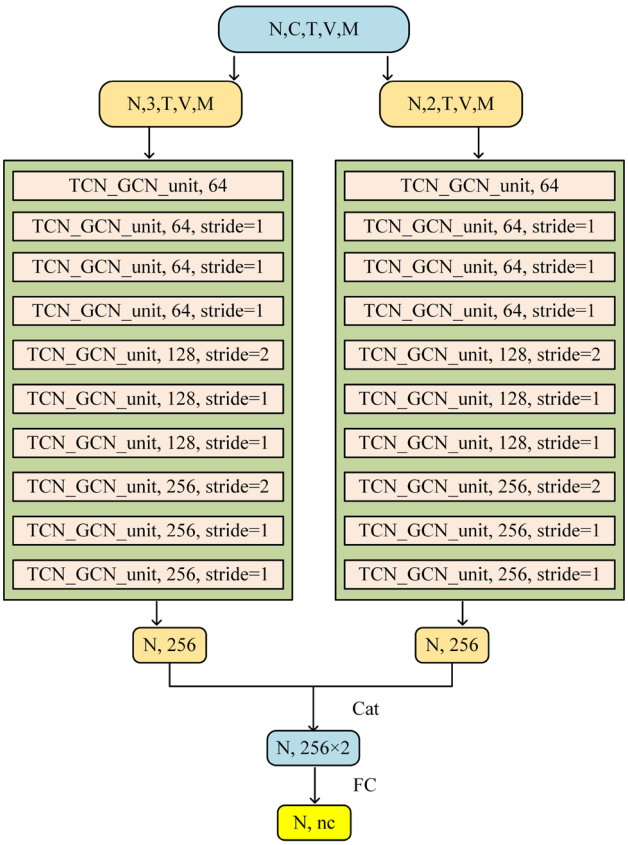


Fig. 7 2s-ShiftGCN network structure diagram

4 Experiment

4.1 Dataset

The OQ35-human and OQ35-keypoint datasets consist of 8760 thermal infrared images captured in various scenes and weather conditions using Hikvision OQ35 cameras. Each image contains one to five human bodies. The datasets were annotated using Labelme and COCO-Annotator tools to create thermal infrared target detection and pose estimation datasets. The InfAR-skeleton dataset is created by combining videos from the InfAR dataset with self-recorded thermal infrared light videos capturing relevant actions. In each frame of the video, the human skeleton is extracted by using the pretrained thermal infrared Alphapose model, and a spatial-temporal graph convolution sequence is constructed. The self-made dataset follows specific rules for video recording and captures 13 common dynamic human actions, including fights, hugs, and other physical actions. The dataset covers a range of contexts varying from simple to complex scenarios. As sensory features like eyes and ears play a minor role in thermal infrared action recognition, we exclude the corresponding nodes, resulting in a skeletal structure represented by 14 nodes. Additionally, supplemental samples of relevant behavioral thermal infrared light videos are added to the original dataset. Each sample in this dataset consists of 30 frames of skeletal sequences, resulting in a total of 32,083 samples. The dataset is divided into a training set and a test set with an 8 to 2 ratio for model training and evaluation purposes. The filtered kinetics-skeleton dataset is a subset of

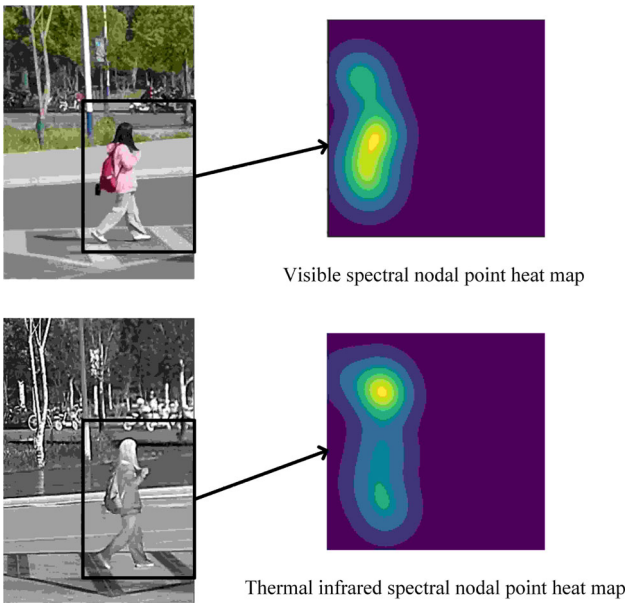


Fig. 8 Visible spectrum and thermal infrared spectrum joint point heat map

the larger kinetics-skeleton dataset. It includes 30 selected actions with high limb correlation, such as clapping and jogging. The dataset has 14 key nodes and comprises 18,877 training samples and 1487 test samples.

4.2 Experimental platform and evaluation criteria

Our experiments were conducted on a platform with an RTX3060 graphics processor, and all models were implemented and trained using the PyTorch framework.

In our experiments,we use five evaluation metrics for target detection algorithms: precision (P), recall (R), mean average precision at IoU with a threshold of 0.50 ($mAP@.5$), mean average precision at different IoU thresholds (ranging from 0.50 to 0.95, in increments of 0.05) denoted as $mAP@.5:.95$, and frames per second (FPS) as an indicator of algorithmic efficiency. These steps guarantee a thorough and impartial comparison of the examined algorithms.

We assess the networks' classification accuracy using TOP-1 and TOP-5 measures. TOP-1 checks if the correct label has the highest classification probability, while TOP-5 considers the top 5 probabilities and checks if the correct label is among them. Besides, we utilize a confusion matrix to assess the effectiveness of each algorithm in action recognition. For judging the recognition accuracy of algorithms for different actions, we consider three evaluation criteria: Accuracy (Acc), Precision (P), and Recall (R) as shown in Table 1.

4.3 Experimental results and analysis

In this subsection, we present a thorough analysis of limitations observed in current models for human detection and pose estimation in thermal infrared environments. We acquired two sets of images namely the visible image set A and the thermal infrared image set B, simultaneously and synchronously in the same scene using a Hikvision OQ35 thermal imager and a visible camera. Each set contains 58 images, ensuring a well-matched dataset for our analysis. To ensure experimental fairness, we used a Hikvision Microimage OQ35 thermal imaging camera with the same focal length and resolution as a regular camera. This ensured a consistent setup and minimized any potential biases in the comparison between the thermal imaging and regular imaging results. We extracted the skeletons of Groups A and B using the YOLOv7 and Alphapose models, trained on the COCO dataset and designed for visible spectral datasets. These skeletons' articulated joints were then superimposed onto the heat map, as illustrated in Fig. 8. This approach enabled a comprehensive comparison and analysis of the models' performance in the thermal infrared environment.

The cosine similarity formula is shown in Eq. 4, which calculates the similarity of the heat map, resulting in a cosine similarity of 0.602 for the case of more distant photography. This observation indicates that existing models fails to accurately extract the skeletal features in thermal infrared images.

$$O = \frac{\sum_{r=0}^{2} M_r U_r}{\sum_{r=0}^{2} M_r \sum_{r=0}^{2} U_r} \tag{4}$$

Target detection experiments were conducted using the self-made thermal infrared dataset. Six YOLO algorithms from YOLOv5 to YOLOv7 were used to conduct comparison experiments on the dataset, and each algorithm was trained with 1000 epochs and compared with each other in five evaluation criteria to find the algorithm that can balance detection accuracy and detection speed. The test results are shown in Table 2.

Based on the experimental results, it can be observed that YOLOv7-tiny achieves a slight reduction in detection accuracy, but significantly improves the detection speed. Comparing the lightweight versions of each YOLO model, YOLOv7-tiny stands out with the fastest detection speed and the least number of model parameters. As a result, YOLOv7-tiny is selected as the thermal imaging human detection model due to its ability to maintain high accuracy, fast detection speed, and lightweight characteristics (Fig. 9).

Pose estimation experiments were conducted using the self-made OQ35-keypoint dataset, which has a limited number of samples, resulting in relatively lower accuracy when training Alphapose from scratch. To enhance the thermal infrared pose estimation performance of our model, we uti-

Table 1 Action recognition evaluation criteria

Confusion matrix	Positive example	Counterexamples
Positive example (real label)	True positive sample (TP)	False inverse sample (FN)
Counterexamples (real labels)	False positive sample (FP)	True inverse sample (TN)

Table 2 Action recognition evaluation criteria

YOLO algorithm	P	R	mAP@.5	mAP@.5:.95	FPS	Params (M)
YOLOv5s	0.918	0.864	0.937	0.556	98	13.7
YOLOv5x	0.926	0.881	0.945	0.572	25	166.9
YOLOv6N	0.882	0.841	0.907	0.542	112	36.7
YOLOv6M	0.883	0.850	0.917	0.550	68	71.3
YOLOv7	0.892	0.866	0.931	0.572	42	71.3
YOLOv7-tiny	0.861	0.875	0.921	0.552	139	11.7

lized a pre-trained model from the person_keypoints training set, which had been trained on the COCO dataset in the visible spectrum [36]. Figure 10 shows a detailed comparison between the heat map human skeleton thermogram extracted by our trained model and the actual skeleton heat map. The cosine similarity between the two sets of thermograms was calculated, resulting in a value of 0.949. The comparison enables us to evaluate our model's accuracy and effectiveness in human pose estimation within thermal infrared environments.

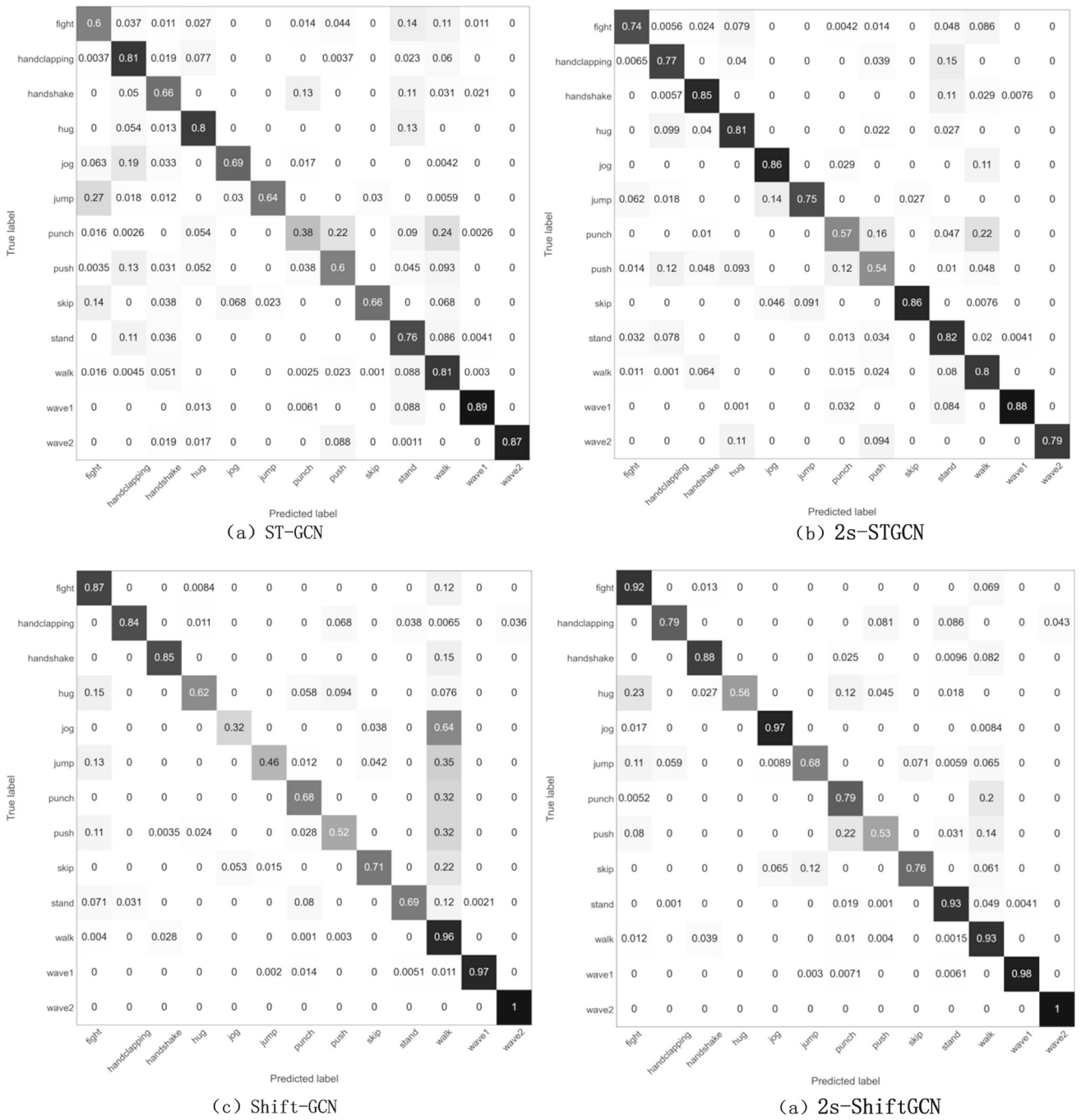
We have introduced variations in noise levels and ambiguity to assess the algorithm's performance. Our additional experimental results encompass the inclusion of Gaussian noise, Poisson noise, periodic noise, Gaussian blur, motion blur, and mean blur. Moreover, we have adjusted the resolution from 720 × 576 to 200 × 200 and 480 × 400 to observe differences. The experimental results are shown in Tables 3 and 4. By conducting cosine similarity calculations on skeleton heatmaps extracted under diverse conditions, we have determined that alterations in resolution, noise levels, or blur conditions do not significantly impact the accuracy of skeleton point extraction, except for periodic noise. Periodic noise, typically induced by equipment malfunctions or power interferences, manifests as periodic fluctuations in image brightness. This noise may introduce noticeable stripes or patterns in the image, thereby affecting its overall quality and visual clarity.

Action recognition experiments were conducted using both the InfAR-skeleton dataset and the filtered visible spectrum dataset kinetics-skeleton. We implemented comparison experiments on two datasets to validate the advantages of 2s-ShiftGCN, training each algorithm for 500 epochs. Table 5 shows the InfAR-skeleton dataset and the Filtered kinetics-skeleton dataset comparison results.

Because the Kinetics-Skeleton dataset contains more types and samples compared to the InfAR-skeleton dataset, Top-1 accuracy and Top-5 accuracy decreased for all four models. Both datasets consist of numerous limb-related actions with large movement amplitudes within short periods. Therefore, owing to its two-stream structure and shift graph convolution, 2s-ShiftGCN maximizes the classification effect, outperforming other networks in action prediction. The model achieves Top-1 and Top-5 accuracies of 88.06% and 98.28% on the InfAR-skeleton dataset, and 55.26% and 83.98% on the filtered kinetics-skeleton dataset, surpassing other networks. The confusion matrices of the four algorithms for recognizing 13 types of actions on the test set of the InfAR-skeleton dataset are shown in Fig. 9. Next, We further analyze the accuracy of the four networks in each act using the confusion matrix and assess the advantages and disadvantages of 2s-ShiftGCN.

According to the confusion matrix, the two-stream structure outperforms the single-stream structure in classifying most motions. Specifically, actions with significant changes captured by a single frame, such as skipping, demonstrated a 12% higher accuracy on Shift-GCN and a 5% higher accuracy on ST-GCN. This improvement can be attributed to the unique feature of capturing motion frames in the two-stream structure. Shift convolution effectively combines spatial shift graph convolution with pointwise convolution, achieving information fusion in spatial and channel dimensions. The non-local shift graph operation enables adaptive learning of joint relationships, resulting in higher performance compared to traditional spatial-temporal graph convolutions [17]. For actions where limb joints are physically distant from each other but their relationship is crucial such as fighting, Shift-GCN outperforms ST-GCN with a 16% and 18% higher accuracy for the single-stream and two-stream structures, respectively. This highlights the effectiveness of shift convolution in capturing important spatial relationships among distant joints.

To assess the performance of 2s-ShiftGCN on different actions, we utilized accuracy (Acc), precision (P), and recall (R) as evaluation criterias. Table 6 presents the results of



(a) ST-GCN (b) 2s-STGCN

(c) Shift-GCN (a) 2s-ShiftGCN

Fig. 9 Confusion matrix of the predicted action of each algorithm

Table 3 The accuracy of skeleton extraction algorithms under various noise and blur conditions

Noise or blur	Gaussian noise	Poisson noise	Periodic noise	Gaussian blur	Motion blur	Mean blur
Cosine similarity	0.905	0.948	0.770	0.946	0.884	0.931

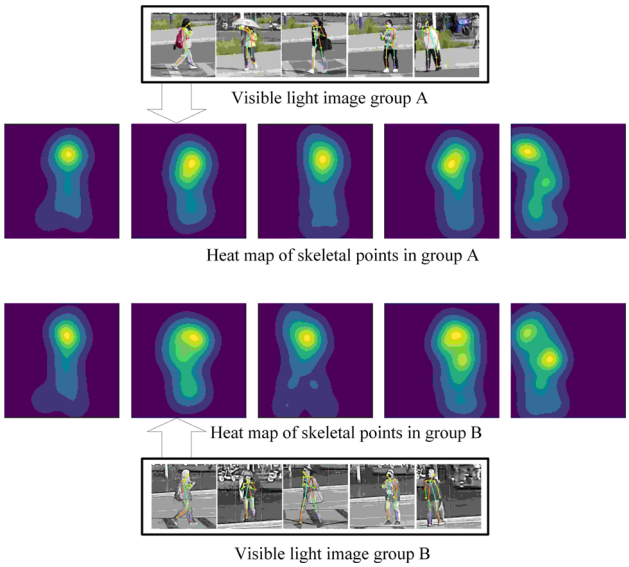


Fig. 10 Heat map A, B comparison

Table 4 The accuracy of algorithm extraction of skeleton at different resolutions

Noise or blur	200 × 200	480 × 400	720 × 576
Cosine similarity	0.896	0.946	0.949

Table 5 Action recognition evaluation criteria

Project	InfAR-skeleton		Filtered kinetics-skeleton	
	Top-1	Top-5	Top-1	Top-5
ST-GCN	0.7659	0.9603	0.4433	0.7653
2s-STGCN	0.7892	0.9679	0.4626	0.7766
Shift-GCN	0.8291	0.9823	0.5084	0.7969
2s-ShiftGCN	0.8806	0.9828	0.5526	0.8398

2s-ShiftGCN in detecting various actions. From the tables, we observed that 2s-ShiftGCN achieved high accuracy rates for single-player actions such as jogging, standing, walking, and waving. However, when it comes to some multi-player overlapping actions, like hugging and pushing, the accuracy drops to 56% and 53%, respectively. This disparity can be attributed to the top-down Alphapose algorithm, which might confuse two-player nodes in cases of high overlap, resulting in errors in skeleton modeling and affecting the training of 2s-ShiftGCN.

Furthermore, we noticed that the precision and recall rates for actions like jumping were relatively low at both 68%. Upon analysis, we attributed this outcome to the nature of jumping, which involves large limb swings that obscure some joint points by the body. This leads to inaccuracies in the skeleton model, subsequently affecting the training of

Table 6 Results of 2s-ShiftGCN detection of different actions

Action	Acc	P	R
Fight	0.9500	0.6695	0.9182
Handclapping	0.7900	0.9294	0.7900
Handshake	0.8800	0.9176	0.8830
Hug	0.5600	1.0000	0.5600
Jog	0.9700	0.9292	0.9745
Jump	0.6800	0.8468	0.6801
Punch	0.7900	0.6633	0.7938
Push	0.5300	0.8018	0.5295
Skip	0.7600	0.9146	0.7555
Stand	0.9300	0.8547	0.9262
Walk	0.9300	0.5797	0.9333
One-handed waving	0.9800	0.9958	0.9837
Two-handed waving	1.0000	1.0000	0.9588

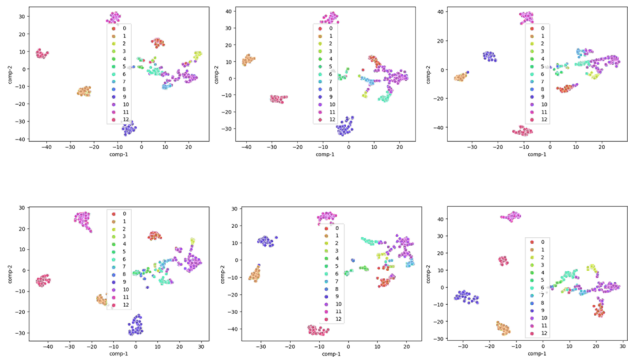


Fig. 11 Output characteristics T-SNE diagram

2s-ShiftGCN and causing the model to misjudge such movements.

In general, we can view neural networks as black box models, in which the information of hidden layers are usually unreadable by users. At this point, the information of the hidden layer needs to be intercepted and downscaled for visualization. T-SNE is a machine learning technique used for dimensionality reduction and data visualization. Its main objective is to map high-dimensional data into a lower-dimensional space while maintaining the similarity relationships between data points. T-SNE is a non-linear dimensionality reduction method ideal for data visualization, as it helps us better understand the structure and clustering of data in two or three dimensions, and making graphical and intuitive model assessments. 2s-ShiftGCN model output features T-SNE visualization is shown in Fig. 11.

The T-SNE plots reveal that most actions were well classified, consistent with the model's accuracy (88.06%). However, classes 3, 5, and 7 (hugging, jumping, and pushing) showed lower accuracy (56%, 68%, and 53%), as indicated by their clustering in the T-SNE plots. Moreover, classes 3

and 0, as well as 7 and 6, show convergence in the T-SNE plots, representing a high misidentification rate for hugging and fighting, and pushing and fighting actions, with rates of 23% and 22%, respectively.

5 Conclusion

In this paper, we propose a novel and lightweight skeleton-based action recognition algorithm tailored for thermal infrared environments. The approach combines target detection, pose estimation, and GCN techniques effectively. Prior to behavior recognition, thermal imaging conducts data desensitization, which involves the removal of sensitive information in the data to protect personal privacy. This process includes obscuring location details and eliminating identifiable information. Thermal imaging facial features, such as facial recognition points and contours, ensuring the anonymity of individuals. This approach is vital for safeguarding data privacy in the context of behavior recognition. Additionally, we propose a joint training algorithm for action recognition in thermal imaging, aiming to address the challenges posed by limited types and samples in thermal imaging datasets, thereby enhancing the model's robustness. Besides, we introduce 2s-ShiftGCN for enhancing action recognition accuracy. The non-local shift graph convolution mechanism addresses challenges in the spatio-temporal perception domain. The newly devised two-stream structure adeptly integrates inter-frame features and fuses first-order and second-order information of human motion. These two designs notably boost action recognition performance. The overall action recognition accuracy achieved by 2s-ShiftGCN model on the InfAR-skeleton dataset is 88.06%. On the filtered kinetics-skeleton dataset, the Top-1 accuracy and Top-5 accuracy of 2s-ShiftGCN are 55.26% and 83.98%, respectively. The experimental findings demonstrate that despite the comparatively limited human body information in thermal infrared images compared to visible images, leveraging pose estimation to extract skeletons enables successful training and prediction of action recognition in both thermal infrared and visible videos.

The subsequent phase could focus on meeting the demands of diverse environments, various individuals, and multiple scenarios during the sample acquisition process. The objective is to develop a comprehensive and diverse thermal infrared pose estimation dataset that caters to various conditions. This effort aims to improve the accuracy and robustness of the thermal infrared pose estimation model, ensuring its effectiveness in various scenarios and applications.

Author Contributions All authors contributed to the study's conception and design. Experimentation and ablation studies were performed by JL, WH, and JW. Data analysis and review were conducted by DH, RH and DT. The first draft of the manuscript was written by JL. Review and editing were performed by HW, and all authors commented on previous versions of the manuscript. The project supervision is done by F. All authors read and approved the final manuscript.

Funding No funding was received for conducting this study.

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code Availability The code that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest There is no conflict of interest in this paper.

References

1. Raza, M.A., Fisher, R.B.: Vision-based approach to assess performance levels while eating. Mach. Vis. Appl. **34**(6), 124 (2023)
2. Gammulle, H., Ahmedt-Aristizabal, D., Denman, S., Tychsen-Smith, L., Petersson, L., Fookes, C.: Continuous human action recognition for human–machine interaction: a review. ACM Comput. Surv. **55**, 1–38 (2022)
3. Gao, C., Du, Y., Liu, J., Lv, J., Yang, L., Meng, D., Hauptmann, A.: Infar dataset: infrared action recognition at different times. Neurocomputing **212**, 36–47 (2016)
4. Jiang, Z., Rozgic, V., Adali, S.: Learning spatiotemporal features for infrared action recognition with 3d convolutional neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 309–317 (2017)
5. Liu, Y., Lu, Z., Li, J., Yang, T., Yao, C.: Global temporal representation based cnns for infrared action recognition. IEEE Signal Process. Lett. **25**, 848–852 (2018)
6. Wang, L., Gao, C., Zhao, Y., Song, T., Feng, Q.: Infrared and visible image registration using transformer adversarial network. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 1248–1252 (2018)
7. Chen, X., Gao, C., Li, C., Yang, Y., Meng, D.: Infrared action detection in the dark via cross-stream attention mechanism. IEEE Trans. Multimed. **24**, 288–300 (2021)
8. Wang, C.-Y., Bochkovskiy, A., Liao, H.: Yolov7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7464–7475 (2022)
9. Fang, H., Xie, S., Tai, Y.-W., Lu, C.: Rmpe: regional multi-person pose estimation. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2353–2362 (2016)
10. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. ArXiv, pp. 7444–7452 (2018)
11. Zhang, X., Demiris, Y.: Visible and infrared image fusion using deep learning. IEEE Trans. Pattern Anal. Mach. Intell. **45**, 10535–10554 (2023)
12. Si, T., He, F., Li, P., Gao, X.: Tri-modality consistency optimization with heterogeneous augmented images for visible-infrared person re-identification. Neurocomputing **523**, 170–181 (2023)
13. Liu, D., Yang, H., Shao, Y.: Fusion of infrared and visible light images for object detection based on CNN. In: 2021 10th International Conference on Internet Computing for Science and Engineering, pp. 110–115 (2021)

14. Guo, H., Tang, T., Luo, G., Chen, R., Lu, Y., Wen, L.: Multi-domain pose network for multi-person pose estimation and tracking. ArXiv, pp. 209–216 (2018)
15. Torralba, A., Russell, B.C., Yuen, J.: Labelme: online image annotation and applications. Proc. IEEE **98**, 1467–1484 (2010)
16. Stefanics, D., Fox, M.: Coco annotator. ACM SIGMultimed. Rec. **13**, 1–1 (2021)
17. Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H.: Skeleton-based action recognition with shift graph convolutional network. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 180–189 (2020)
18. Ramasinghe, S., Rodrigo, R.: Action recognition by single stream convolutional neural networks: an approach using combined motion and static information. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), pp. 101–105 (2015)
19. Jain, M., Jégou, H., Bouthemy, P.: Better exploiting motion for better action recognition. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2555–2562 (2013)
20. Qi, D., Su, L., Song, J., Cui, E., Bharti, T., Sacheti, A.: Imagebert: cross-modal pre-training with large-scale weak-supervised image-text data. arXiv:2001.07966 (2020)
21. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724–4733 (2017)
22. Ji, K., Lei, W., Zhang, W.: A deep retinex network for underwater low-light image enhancement. Mach. Vis. Appl. **34**(6), 122 (2023)
23. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1725–1732 (2014)
24. Donahue, J., Hendricks, L.A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2625–2634 (2014)
25. Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 4489–4497 (2014)
26. Zhou, Y., Sun, X., Zha, Z., Zeng, W.: Mict: mixed 3d/2d convolutional tube for human action recognition. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 449–458 (2018)
27. Zhu, X., Wang, Y., Dai, J., Yuan, L., Wei, Y.: Flow-guided feature aggregation for video object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 408–417 (2017)
28. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. arXiv:1406.2199 (2014)
29. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Gool, L.: Temporal segment networks: towards good practices for deep action recognition. ArXiv, pp. 20–36 (2016)
30. Zhu, Y., Lan, Z., Newsam, S., Hauptmann, A.: Hidden two-stream convolutional networks for action recognition. arXiv:1704.00389 (2017)
31. Liu, K., Liu, W., Gan, C., Tan, M., Ma, H.: T-c3d: temporal convolutional 3d network for real-time action recognition. ArXiv, pp. 7138–7145 (2018)
32. Zhang, X., Zeng, H., Guo, S., Zhang, L.: Efficient long-range attention network for image super-resolution. ArXiv, pp. 649–667 (2022)
33. Tan, M., Le, Q.V.: Efficientnet: rethinking model scaling for convolutional neural networks. arXiv:1905.11946 (2019)
34. Zhang, G., Zhu, Y., Wang, H., Chen, Y., Wu, G., Wang, L.: Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5682–5692 (2023)
35. Tsai, D.-M., Chiu, W.-Y., Lee, M.-H.: Optical flow-motion history image (OF-MHI) for action recognition. Signal Image Video Process. **9**, 1897–1906 (2015)
36. Papandreou, G., Zhu, T.L., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3711–3719 (2017)

Jishi Liu was born in Shaoyang City, Hunan Province, China, in 1999. He received his bachelor's degree in Electronic Information Science and Technology in 2021. Currently, he is pursuing his master's degree in IC at Hunan University of Science and Technology, Hunan, China. His research interests include image detection, edge computing and and deep learning based applications.

Huanyu Wang was born in Jinchang City, Gansu Province, China, in 1995. He received the Ph.D. and M.S. degree in Information and Communication Technology from KTH Royal Institute of Technology, Stockholm, Sweden, in 2023 and the B.S. degree in Electronic Information Engineering from Dalian University of Technology, Dalian, China. He is currently an Assistant Professor at the School of Computer Science and Engineering at Hunan University of Science and Technology. His current research interest includes hardware security, side-channel analysis and deep learning based applications.

Junnian Wang received the bachelor's degree from the Department of Modern Physics, Lanzhou University, in 1991, the master's degree in radio physics from the School of Information Science and Engineering, Lanzhou University, in 2000, and the Ph.D. degree in control theory and control engineering from the School of Information Science and Engineering, Central South University, in 2006. He has undertaken four projects of the National Natural Science Foundation of China and more than ten other provincial and ministerial level research projects. He has published more than 50 scientific papers, including more than 20 SCI/EI papers. His research interests include deep learning, intelligent information processing, and fault diagnosis. He received the Second Prize of Hunan Provincial Science and Technology Progress Award.

Dalin He was born in Changsha City, Hunan Province, China, in 2000. He received the bachelor's degree in Optoelectronic Information Science and Engineering from the Hunan University of Science and Technology, Hunan, China, in 2022, where he is currently pursuing the master's degree in IC. His research interests include hardware encryption, side channel analysis, and maliciouos code intrusion detection.

Ruihan Xu was born in Tongliao City, Neimenggu Province, China, in 1998. He received a bachelor's degree in Electronic Information Science and Technology from Hunan University of Science and Technology in 2019. He obtained a master's degree in electronic information from Hunan University of Science and Technology in 2023. Currently, he is working on Automotive Steering Control Algorithms at BYD Automobile Industry Co., Ltd.

Xiongfeng Tang was born in Chongqing City, China, in 1998. He received the bachelor's degree in Optoelectronic Information Science and Engineering from the Hunan University of Science and Technology, Hunan, China, in 2021, where he is currently pursuing the master's degree in IC. His research interests include power load decomposition.