PAPER

2D Human Skeleton Action Recognition Based on Depth Estimation

Lei WANG[†,††], ***Member***, **Shanmin YANG**[†††], **Jianwei ZHANG**[†], ***and*** **Song GU**[††a)], ***Nonmembers***

SUMMARY Human action recognition (HAR) exhibits limited accuracy in video surveillance due to the 2D information captured with monocular cameras. To address the problem, a depth estimation-based human skeleton action recognition method (SARDE) is proposed in this study, with the aim of transforming 2D human action data into 3D format to dig hidden action clues in the 2D data. SARDE comprises two tasks, i.e., human skeleton action recognition and monocular depth estimation. The two tasks are integrated in a multi-task manner in end-to-end training to comprehensively utilize the correlation between action recognition and depth estimation by sharing parameters to learn the depth features effectively for human action recognition. In this study, graph-structured networks with inception blocks and skip connections are investigated for depth estimation. The experimental results verify the effectiveness and superiority of the proposed method in skeleton action recognition that the method reaches state-of-the-art on the datasets.

key words: *action recognition, depth estimation, muti-tasks learning, graph structure, video surveillance*

1. Introduction

The development of video surveillance is manifested by the automatic identification of the human actions in the scene [1]. However, surveillance videos are primarily captured with 2D cameras, causing low accuracy and efficiency in HAR [2]. Although the accuracy can be increased using multiple cameras from multiple perspectives, the cost and operation are expensive. Wearing sensors or calibration points [3] to extract human skeleton data for action recognition cannot be achieved with passive personnel. The 2D joint positions of the skeleton cannot fully represent accurate action information since certain motion ambiguity is triggered. At the same time, 3D data covers more information and shows more significant advantages over 2D data. Therefore, existing research has suggested that HAR tasks are capable of achieving higher accuracy on 3D data rather than on 2D data. Thus, the relevant depth information of human action should be obtained from 2D data for HAR in most video surveillance application scenarios, such that it has become a challenging technology and current research hotspot, which is the primary purpose of this paper, focusing on the methods of transforming 2D human action data into 3D format and then to conduct the human action recognition task based on it to better enhance the recognition accuracy from the perspective of data dimensional improvement.

The HAR methods based on video images are notably influenced by environmental factors such as lighting and occlusion. Moreover, the large amount of image data leads to complex calculations. In contrast, methods based on human skeletons have more advantages [4], that the skeleton can highly summarize the motion characteristics of the human body and represent the spatial connections between body joints to improve the accuracy, calculation speed, and stability of HAR.

Depth estimation can fall into binocular and monocular depth estimation. The binocular method employs binocular disparity to estimate depth, and the monocular method primarily conforms to geometric perspective relationships, occlusion relationships, focusing situations, and image colour textures. This study focuses on monocular depth estimation since most surveillance videos are captured with monocular cameras. Monocular depth estimation networks integrate the above clues to develop high-level semantic features and reason the depth information.

To fully explore the actional clues of the human skeleton in 2D surveillance videos at a higher level for action recognition, a monocular depth estimation-based human skeleton action recognition method (SARDE) is proposed in this paper, as in Fig. 1. The SARDE comprises two tasks: (1) human skeleton action recognition and (2) monocular depth estimation. The two tasks are integrated in an end-to-end training manner to fully utilize the correlation between action recognition and depth estimation, sharing parameters to learn better the human action features for HAR. By utilizing the constraints in the two tasks by multi-task learning, the model performance can be further enhanced. Moreover, the accuracy of human skeleton action recognition can be increased. The main contributions of this study are as follows.

- The transformation of 2D human skeleton action data into 3D format is investigated to dig for hidden action clues in the 2D data to better enhance the recognition accuracy from the perspective of data dimension improvement;
- The depth estimation methods of 2D, graph-structured skeleton data are innovatively designed and verified with deep-learning neural networks;
- A multi-task approach is adopted in an end-to-end manner to comprehensively exploit the constraints between

Manuscript received October 27, 2023.
Manuscript revised January 9, 2024.
Manuscript publicized February 27, 2024.

[†] The authors are with Sichuan University, China.
[††] The authors are with Chengdu Aeronautic Polytechnic, China.
[†††] The author is with Chengdu University of Information Technology, China.
a) E-mail: gusong@cap.edu.cn

DOI: 10.1587/transinf.2023EDP7223

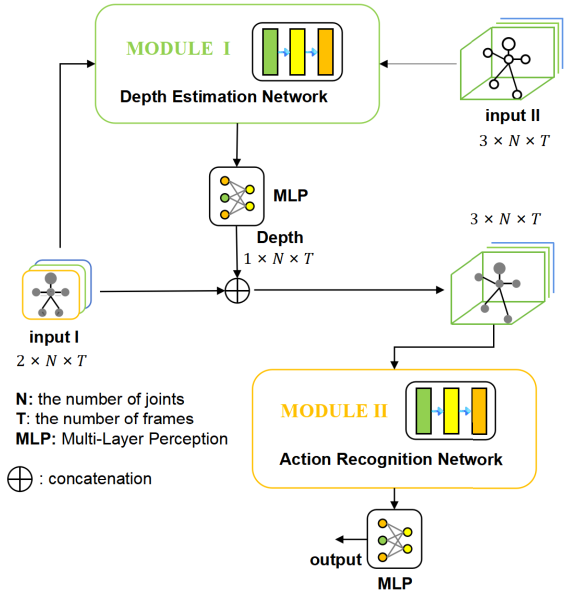


Fig. 1 The framework of SARDE. Module I inputs the 2D skeleton joints to output the estimated depth. For its training, the 1D depth information in the 3D dataset (NTU-RGB+D) is the ground truth. Then, the 3D reconstruction of the action data is accomplished by concatenating the 2D input and the estimated 1D depth. Module II utilizes the 3D data for action recognition.

action recognition and depth estimation tasks while the training parameters are shared among them. Furthermore, a combined loss function is proposed for the multi-task learning to balance the independence and correlation between tasks to enhance the model performance.

2. Related Works

2.1 Video Depth Estimation

Deep learning methods design the structure and training loss of the neural networks. D. Eigen et al. [5] introduced multi-scale networks in a supervised manner for monocular depth estimation. Laina et al. [6] presented residual learning, exhibiting a deeper network and the capability of outputting higher-resolution depth maps. Godard et al. [7] transformed the depth estimation problem into stereo-matching between left and right views. Wang et al. [8] proposed a self-supervised correction mechanism capable of learning the internal structure of human action from 2D images. Sun et al. [9] proposed a structure-aware regression method that employs re-parameterized joint features to replace original features. Literature [10] uses dilated convolution to expand the perception area of the model to acquire richer global information. Ranftl et al. [11] used Transformers as feature extractors. Wang et al. [12] proposed an embedded loss to effectively measure the semantic spatial distance between depth estimation results and actual values. DenseNet [13] has also been extensively employed in deep estimation networks. Some research [14] considered depth estimation as an image generation problem, using generative adversarial networks to predict depth maps. Jiao et al. [15] combined depth estimation and semantic segmentation tasks.

2.2 Human Skeleton Action Recognition

Many studies utilized CNN (Convolutional Neural Network) and RNN (Recurrent Neural Network) to determine more impressive features for skeleton HAR. With the use of GCN (Graph Convolutional Network) and LSTM (Long Short Term Memory), Cho et al. [16] introduced skeleton motion history for the recognition task. Wang and Wang [17] developed a two-stream RNN network to abstract the action features. Si et al. [18] used graph convolution operators in LSTM to determine the skeleton features. Moreover, the spatiotemporal relations were captured using a graph convolutional LSTM network. Attention mechanisms and self-supervised methods have been progressively employed in recent years. Liu et al. [19] aggregated long-range spatial features and learned long-distance temporal correlations by introducing a large kernel attention operator. Zhang et al. [20] proposed MixSTE with temporal and spatial transformer blocks to learn joint correlations. SATD-GCN [21] utilized spatial attention pooling and temporal graph convolution to acquire fine-grained information for action recognition. Cho et al. [22] proposed three variants of self-attention networks. Tu et al. [23] introduced a semi-supervised modality for skeleton-based action recognition. The X-CAR [24] aimed to increase the performance in the semi-supervised scenario through an adaptive-combination augmentation mechanism. Su et al. [25] employed an encoder-decoder RNN to achieve semi-supervised learning. Sachin and Subrahmanyam [26] invented image depth estimation based on CNN for HAR, but they did not investigate the depth estimation methods applicable to graph-structured skeleton action data.

As indicated by the above research, existing research on skeleton HAR has primarily focused on mining high and deep-level action features for recognition from the perspective of designing network structures. Nevertheless, the principles of network construction have yet to be investigated. Since a considerable amount of excellent research on image depth estimation can be referenced, they can be transferred to depth estimation for graph-structured skeleton action data with specific strategies.

3. SARDE Framework

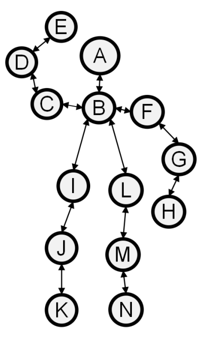
The SARDE comprises two modules, Fig. 1, which correspond to the two tasks of it. Module I estimates the depth of graph-structured skeleton joints based on GCN (Graph Convolutional Network). It takes the 2D coordinates of the joints in the image as the input and outputs the estimated depth. Given a set of skeleton joints $S_I \in \mathbb{R}^{2\times N\times T}$, extracted by OpenPose [27] from video images, module I aims to estimate the joint depth D_{pre} as close as possible to the ground truth D_{target} in supervised learning manner. Module II employs the 2D image coordinates of the joints S_I combined with the estimated depth D_{pre} as the input $S_{II} \in \mathbb{R}^{3\times N\times T}$ to output predicted action class C_{pre} as close as possible to the ground truth C_{target} with supervised learning. N represents the number of joints in a static skeleton, and T is the number of frames in the action video.

3.1 Depth Estimation Network

Depth estimation is typically applied in image and video scenarios, in which the action information is organized by 2D-pixel coordinates of the image, and many of the above-described methods are based on the properties of actions in the image. However, the skeleton joints are not organized in this manner. The most proper way for them is graph structure rather than image pixels. Therefore, the conventional depth estimation methods will no longer be applicable for skeleton joint situations. A new approach of GCN-based depth estimation is adopted to address this issue.

Human actions can be considered a series of joint graph frames $G = (V, E)$, where the joints and their connections are organized as nodes and edges, V presents the nodes, E are the edges (e.g., the adjacent matrix A), revealing the connection and significance of neighbour nodes to the root node. $\mathcal{A}_i = \{d(i,j) < Th\}$ is the neighbour set of node i with threshold Th; $d(i,j)$ is the distance from node i to j, Fig. 2.

The GCN-based approach refers to an extension of CNN in a graph structure, exhibiting the capability of modelling graph-based skeletal data [28] and effectively representing the geometric and structural correlation between nodes. The GCN operator [29] is expressed as Eq. (1), X_{out} denotes the graph convolutional output of a node, $GCN()$ is the GCN operator, the labelling function $label()$ assigns labels to nodes in the neighbour set, such that different significance is implemented in the neighbour nodes, with the weight function $W()$, Z_i is for normalization. Assume the input is X_{in} and the degree matrix $\Lambda^{ii} = \sum_j A^{ij}$, the GCN operator cab be Eq. (2).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	1	1	0	0	0	0	0	0	0	0	0	0	0	0
B	1	1	1	0	0	1	0	0	1	0	0	1	0	0
C	0	1	1	1	0	0	0	0	0	0	0	0	0	0
D	0	0	1	1	1	0	0	0	0	0	0	0	0	0
E	0	0	0	1	1	0	0	0	0	0	0	0	0	0
F	0	1	0	0	0	1	1	0	0	0	0	0	0	0
G	0	0	0	0	0	1	1	1	0	0	0	0	0	0
H	0	0	0	0	0	0	1	1	0	0	0	0	0	0
I	0	1	0	0	0	0	0	0	1	1	0	0	0	0
J	0	0	0	0	0	0	0	0	1	1	1	0	0	0
K	0	0	0	0	0	0	0	0	0	1	1	0	0	0
L	0	1	0	0	0	0	0	0	0	0	0	1	1	0
M	0	0	0	0	0	0	0	0	0	0	0	1	1	1
N	0	0	0	0	0	0	0	0	0	0	0	0	1	1

Fig. 2 The skeleton graph structure. The left figure is the natural connection of the human skeleton joints (A to N) in one graph frame, lots of these frames are organized consecutively in time sequence as an action series. The right figure is the adjacency matrix A of one frame. As indicated by 0, 1 in A, the corresponding joints are connected or disconnected.

$$X_{out}(i) = \sum_{j \in \mathcal{A}_i} \frac{1}{Z_i(j)} GCN(j) \cdot W(label(j)) \tag{1}$$

$$X_{out} = \sum \Lambda^{-\frac{1}{2}} A \Lambda^{-\frac{1}{2}} X_{in} W \tag{2}$$

Two depth estimation networks are designed based on the current famous network structures to verify the effectiveness of our methods; one is the series network, Fig. 3, and the other is the inception network, Fig. 4. To be specific, the

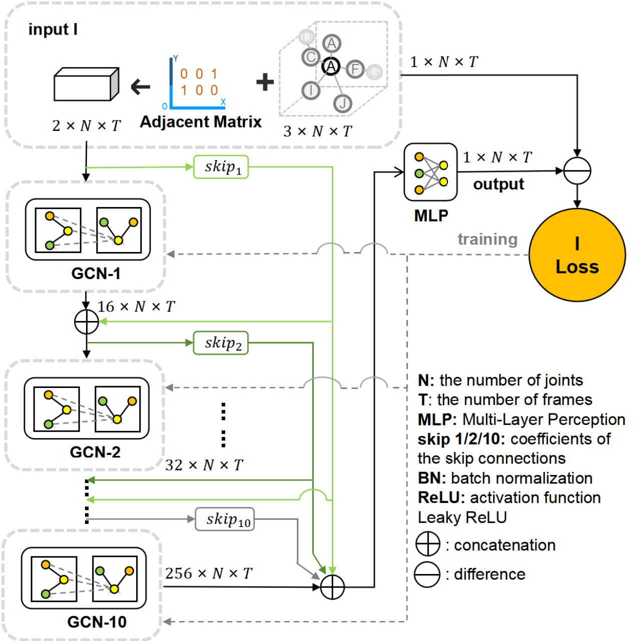


Fig. 3 The series network for depth estimation. There are 10 GCN blocks connected in series with sip connections. The number of data channels increases with GCN operations to obtain high-level action representations gradually. BN and ReLU perform regularization and nonlinear operations on skeleton data; the MLP regresses the estimated depth.

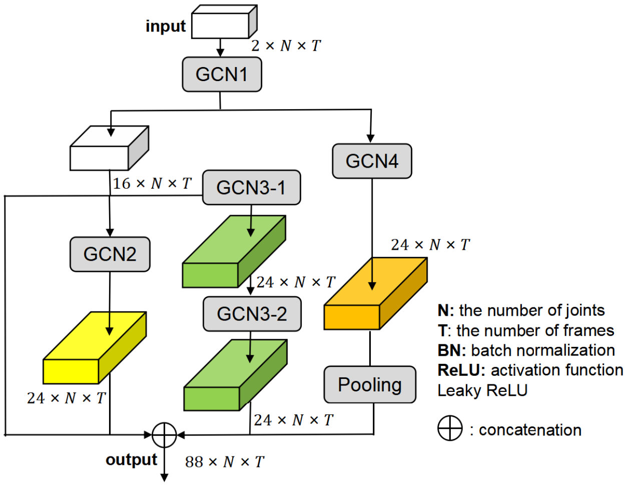


Fig. 4 The inception block for depth estimation. Four parallel sub-networks exist in the block. The channel number and structure of the sub-networks are well designed to achieve the feature extraction capability of the human skeleton.

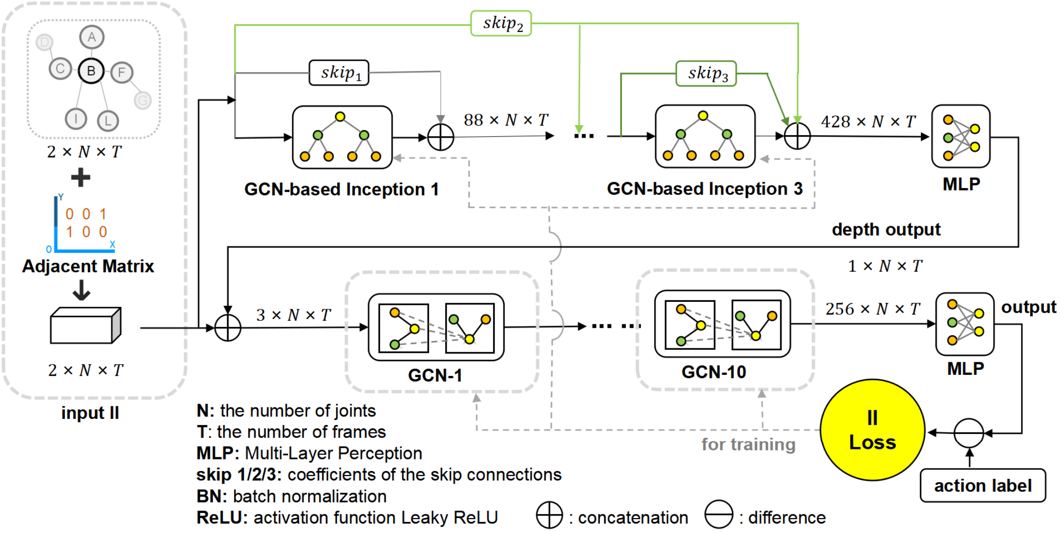


Fig. 5 The action recognition network. The 2D action data, in conjunction with the adjacent matrix, are fed into the depth estimation network to output the estimated depth of the skeleton joints, which is concatenated with itself to form the 3D data for action recognition. Then, the action recognition loss can be calculated to encourage the network training, such that the GCN and MLP blocks can be optimized while the values of the skips are recognized as hyper-parameters.

3D dataset is utilized for the training in the way that the first two dimensions of the data, in conjunction with the adjacent matrix, serve as the input to output the estimated 1-D depth information, and the last one dimension serves as the ground truth. Accordingly, the depth estimation loss can be calculated to encourage network training, optimizing the GCN and MLP blocks. Inspired by the DenseNet [30], skip connections are added among the convolution blocks to enhance the generalization property, whereas the values of the skips are hyper-parameters. Conventional series connection methods achieve feature extraction by increasing the number of convolutional layers and channels, with the disadvantages of having too many parameters, limited generalization, and vanishing gradients. The inception [31] block expands the width of the network by replacing the serial convolution with parallel convolution to overcome the defects. It is introduced and then modified in the inception network to combine with GCN operation for exploiting feature extraction and depth estimation capabilities of the human skeleton. The dimensions of the GCN features are listed in Table 1.

3.2 Action Recognition Network

Figure 5 depicts the structure of the action recognition network to predict human action as close as possible to the ground truth with supervised learning. The features are mainly extracted by the series-stacked GCN blocks, and the classification of action is achieved by the MLP operation. The 3D data is concatenated by the 2D input and the estimated 1D depth to be served as the network input, utilizing the 3D action data to achieve higher accuracy of HAR.

Table 1 The dimensions of the GCN features. In the series network, the channel of the graph features increases gradually with its forward propagation. In the Inception block, the graph features are processed in parallel with different channel numbers and then concatenated.

	No.	Input Size	Output Size	Kernel Size
Series	1	2 × 2 × 150 × 18	2 × 64 × 150 × 18	1 × 1
	2	2 × 64 × 150 × 18	2 × 64 × 150 × 18	1 × 1
	3	2 × 64 × 150 × 18	2 × 64 × 150 × 18	1 × 1
	4	2 × 64 × 150 × 18	2 × 128 × 150 × 18	1 × 1
	5	2 × 128 × 150 × 18	2 × 128 × 150 × 18	1 × 1
	6	2 × 128 × 150 × 18	2 × 128 × 150 × 18	1 × 1
	7	2 × 128 × 150 × 18	2 × 128 × 150 × 18	1 × 1
	8	2 × 128 × 150 × 18	2 × 256 × 150 × 18	1 × 1
	9	2 × 256 × 150 × 18	2 × 256 × 150 × 18	1 × 1
	10	2 × 256 × 150 × 18	2 × 256 × 150 × 18	1 × 1
Inception	GCN1	2 × 2 × 150 × 18	2 × 16 × 150 × 18	1 × 1
	GCN2	2 × 16 × 150 × 18	2 × 24 × 150 × 18	1 × 1
	GCN3-1	2 × 16 × 150 × 18	2 × 24 × 150 × 18	1 × 1
	GCN3-2	2 × 24 × 150 × 18	2 × 24 × 150 × 18	1 × 1
	GCN4	2 × 2 × 150 × 18	2 × 24 × 150 × 18	1 × 1
	Pooling	2 × 24 × 150 × 18	2 × 24 × 150 × 18	1 × 1

3.3 Training Loss

The multi-task training loss $\mathcal{L}$ comprises two parts, Eq. (3), i.e., the depth estimation loss $\mathcal{L}_I$ and the action recognition loss $\mathcal{L}_{II}$, corresponding to the two tasks above. α is the ratio to adjust the two losses.

$$\mathcal{L} = \alpha\mathcal{L}_I + (1-\alpha)\mathcal{L}_{II} \tag{3}$$

The depth estimation refers to a regression issue, generally employing the MSE (mean square error) as the loss function. However, the MSE is sensitive to outliers and easily dominated by large values. Since the depth relationship among joints will not change due to global scale variation, the scale-invariant MSE [5] is adopted and modified to calculate the depth estimation loss, which is calculated in the logarithmic domain, exploiting the scale invariance of the depth values

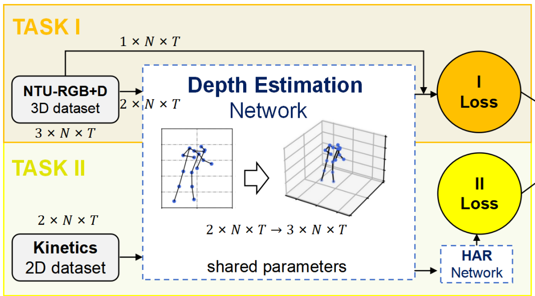


Fig. 6 The training loss of the multi-task SARDE. The NTU-RGB+D 3D dataset is adopted to train the depth estimation network. The Kinetics 2D dataset serves as the input data for human action recognition (HAR) after it has been 3D reconstructed with the depth estimation network. Loss I $\mathcal{L}_I$ and loss II $\mathcal{L}_{II}$ correspond to the two tasks of the model.

to avoid vague estimation caused by significant differences in depth. λ is the factor to adjust the ratio between MSE and scale-invariant MSE to balance their influences, improving the scale perception ability, Eqs. (4) and (5).

$$\mathcal{L}_I = \frac{1}{|N|}\sum_{i\in X_{II}} R(i)^2 - \frac{\lambda}{N^2}\left(\sum_{i\in X_{II}} R(i)\right)^2 \qquad (4)$$

$$R(i) = \lg D_{pre}(i) - \lg D_{target}(i) \qquad (5)$$

The action recognition is a classification issue, which usually uses the cross-entropy in Eq. (6) as the loss function. Take the classification with the highest probability in Eq. (7) as a result to calculate the cross-entropy loss, encouraging the network to be trained progressively to predict the target class with the labelled ground truth. s_n denotes the predicted score of the action belonging to the n^{th} class in the dataset, K represents the number of the class, and c_n is the predicted probability of belonging to the n^{th} class.

$$\mathcal{L}_{II} = -\sum_{u=1}^{K} c_u^{pre} \lg c_u^{target} - \sum_{v=1}^{K} c_v^{target} \lg c_v^{pre} \qquad (6)$$

$$c_n = \frac{e^{s_n}}{\sum_{m=1}^{K} e^{s_m}} \qquad (7)$$

3.4 Multi-Task Constraints

As elaborated in Fig. 6, the two tasks of depth estimation and action recognition in the multi-task approach are trained simultaneously with different 2D and 3D datasets. They would both be fed into the depth estimation network for 3D reconstruction that shares the same parameters to effectively exploit the constraints between the two tasks with losses $\mathcal{L}_I$ and $\mathcal{L}_{II}$. The two losses are integrated into one total loss $\mathcal{L}$ to conduct end-to-end training for recognition accuracy improvement.

4. Experiment and Discussion

In this section, the performance of the proposed method was evaluated and compared with other algorithms on the Kinetics [32] dataset, and the experiment results strongly demonstrate the efficiency and superiority of the proposed method.

4.1 Datasets and Evaluation Metrics

Kinetics Datasets comprise 300,000 raw video clips belonging to 400 action classes from the internet; 240,000 clips are for training, and the rest are for testing. The raw video clips are transformed into skeleton data frames/clips using the preprocessing methods with joint locations in pixel coordinate format as the inputs. The publicly available OpenPose algorithm is adopted to obtain the 2D coordinates (x, y) and their confidence scores Dc of the joints as the input tuples (x, y, Dc) arranged in an array, and the skeleton data sequence comprises the arrays in time order. The size of the skeleton data is (C, T, N) where C denotes the channel number of the tuple, which is 3 for (x, y, Dc), T represents the number of clips which are going to be padded as 150 or 300 in the experiments, N expresses the number of the joint in the respective clip, reaching 18 and 25 in the experiment.

The classification accuracy Acc is used to evaluate the performance of the proposed method in Eq. (8), $Q_{classified}$ denotes the number of actions that have been classified correctly, Q_{total} represents the action numbers. Specifically, TOP-1 and TOP-5 are recommended as the metrics in Eq. (9), Q_{TOP1} presents the number of actions with ground truth labels included in the 1st classification probability among all the test actions, and TOP-5 shares the exact definition.

$$Acc = \frac{Q_{classified}}{Q_{total}} \qquad (8)$$

$$Acc_{TOP1} = \frac{Q_{TOP1}}{Q_{total}} \qquad (9)$$

4.2 Implementation Details

The experiments were performed with Ubuntu 18.04, Intel (R) Core (TM) i7-7700K CPU @ 4.20 GHz processor with 32 GB memory, NVIDIA GTX 2080Ti graphics card with 10G memory, Python version 3.6.9, as well as Pytorch 1.8.1. The initial learning rate is 0.1 and declines by 0.1 per 50 epochs. The batch size is 16, the epochs range from 350 to 400, and the SGD (stochastic gradient descent) serves as the optimiser in the training. The number of nodes N reaches 18 and 25, $\alpha = 0.97$, $\lambda = 0.5$, $skip = 1$.

The 3D format NTU-RGB+D dataset [33] is employed for the training of the depth estimation network. It has 120 human action classes, containing 114,480 action clips performed by 40 persons from three perspectives simultaneously with 25 body joints in the skeleton. To match the skeleton joints in Kinetics and NTU-RGB+D, the correspondence is established in Table 2.

4.3 Ablation Study

The SARDE was analyzed based on four groups with dif-

Table 2 Correspondence of the skeleton joints in Kinetics and NTU-RGB+D dataset. The joints in Kinetics are extracted and defined with the OpenPose algorithm, while the joints in NTU-RGB+D are captured and defined with Kinetics V2 cameras. The numbers represent the index numbers of the joints in the skeleton structure. To match the joints, 11 of 25 joints in the NTU-RGB+D are deleted, while its head joint corresponds to 5 joints in the Kinetics (i.e., 0, 14, 15, 16, 17), and "-" indicates no correspondence. On that basis, the two datasets share the same adjacent matrix.

NTU-RGB+D	Kinetics	NTU-RGB+D	Kinetics
base of spine (1)	-	left foot (16)	-
middle of spine (2)	-	right hip (17)	R Hip (8)
neck (3)	-	right knee (18)	R Knee (9)
head (4)	Nose (0)	right ankle (19)	R Ankle (10)
left shoulder (5)	L Shoulder (5)	right foot (20)	-
left elbow (6)	L Elbow (6)	spine (21)	Neck (1)
left wrist (7)	L Wist (7)	tip of left hand (22)	-
left hand (8)	-	left thumb (23)	-
right shoulder (9)	R Shoulder (2)	tip of right hand (24)	-
right elbow (10)	R Elbow (3)	right thumb (25)	-
right wrist (11)	R Wist (4)	head (4)	R Eye (14)
right hand (12)	-	head (4)	L Eye (15)
left hip (13)	L Hip (11)	head (4)	R Ear (16)
left knee (14)	L Knee (12)	head (4)	L Ear (17)
left ankle (15)	L Ankle (13)	-	-

Table 3 Evaluation of the scale-invariant MSE loss, GCN, and skip connections. The benchmark with one inception block and no skip connections is performed to verify the effectiveness of the scale-invariant MSE loss versus MSE loss under graph-structured skeleton joints from row 2 to 3; the benchmark with two inception blocks and skip connections is for the ablation results between CNN and GCN structures, from 4 to 5; and the benchmark with ten GCN blocks in series are utilized for the evaluation of the skip connection structures.

Networks	Settings	TOP-1	TOP-5
1 inception no skips	MSE	29.2%	55.0%
	scale-invariant MSE	**33.5%**	**57.7%**
2 inceptions with skips	CNN	28.3%	54.4%
	GCN	**34.8%**	**58.4%**
GCN series	skips	36.7%	60.0%
	no skips	**36.7%**	**60.1%**

Table 4 Comparison between series and inception blocks. It compares the performance of the GCN series and Inception networks with the equivalent module size. The GCN series network adopted 10 GCN blocks with skip connections, while the Inception network has three inception blocks with skip connections.

Networks	TOP-1	TOP-5
3 inceptions with skips	35.8%	58.7%
10 GCNs with skips	**36.7%**	**60.1%**

Table 5 Permance comparisons with State-of-the-Art on 2D Kinetics dataset in terms of the accuracy TOP-1 and TOP-5.

Networks	TOP-1	TOP-5	Year
Temporal Conv [36]	17.5%	39.8%	2018
ST-GCN [37]	30.7%	52.8%	2018
AS-GCN [28]	34.8%	56.5%	2019
1s-AGCN [38]	35.1%	57.1%	2019
4s-DGNN [39]	36.9%	59.6%	2019
1s-AAGCN [40]	36.0%	58.4%	2020
SAN [22]	35.1%	55.7%	2020
MSSF-GCN [41]	37.0%	59.8%	2021
AAM-GCN [42]	37.5%	60.5%	2021
SATD-GCN [21]	36.6%	59.8%	2022
CD-JBF-GCN [23]	36.5%	59.6%	2022
1s-HybridNet [34]	**37.7%**	60.3%	2023
AWD-GCN [35]	37.2%	**61.0%**	2023
LKA-GCN(2s) [19]	37.8%	60.9%	2023
4s STF-Net [43]	36.1%	58.9%	2023
Ours SARDE Inception	35.8%	58.7%	-
Ours SARDE Series	36.7%	60.1%	-

ferent settings to verify the significance of each part in it: Tables 3 and 4. The experiment results show the GCN-based series network without skip connections and trained by scale-invariant loss can achieve better HAR performance.

In Table 3, the recognition accuracy with scale-invariant MSE TOP-1 and TOP-5 are 33.5% and 57.7%, respectively, which are 4.3% and 2.7% higher than that of MSE 29.2% and 55.0%, showing that the scale-invariant MSE loss helps the network to obtain high-level depth clues in graph-structured joints rather than MSE loss. The recognition accuracy with GCN are 34.8% and 58.4%, respectively, which are 6.5% and 4.0% higher than that of CNN 28.3% and 54.4%, showing that the GCN-based convolutions exhibit better convergence characteristics in training than conventional CNNs, and they take on critical significance in increasing the HAR accuracy. The skip connections are usually utilized in very deep neural nets to decrease the vanishing of the gradient. Since the networks in the proposed method are not sufficiently deep, the advantages of skip connections have yet to be indicated from the results. The recognition accuracy with skip connections are 36.7% and 60.0%, respectively, which are 36.7% and 60.0% without skip connections, showing less improvement. However, this method is still employed in our method to enhance the generalization property of the networks.

In Table 4, the recognition accuracy with GCN series are 36.7% and 60.1%, respectively, which are 0.9% and 1.4% higher than that of Inception blocks 35.8% and 58.7%, showing that the GCN series network can perform better than the Inception network with the current network structures and computational power conditions.

4.4 Comparison with State-of-the-Art

Compared with current state-of-the-art skeleton-based HAR methods, shown in Table 5, our proposed method SARDE Series reaches 36.7% in TOP-1 and 60.1% in TOP-5, which are distinctly close to the current highest accuracy, i.e., 37.7% in TOP-1 [34] and 61.0% in TOP-5 [35]. Compared with other algorithms whose accuracies outperform our proposed method this year, in Table 6, the parameters and FLOPs of the proposed SRADE method are relatively smaller in size and lighter, indicating that the SARDE can achieve prominent accuracy with less computational resources. It shows that the depth action clues hidden in 2D action data can be restored and utilized with depth estimation methods to enhance action recognition performance effectively. This is the

Table 6 Performance comparisons with State-of-the-Art (2023) on 2D Kinetics dataset in terms of the accuracy TOP-1, TOP-5, Params and FLOPs.

Networks	TOP-1	TOP-5	Params	FLOPs
			Million	Giga
1s-HybridNet [34]	37.7%	60.3%	8.3	40.5
AWD-GCN [35]	37.2%	61.0%	10.6	63.8
LKA-GCN(2s) [19]	37.8%	60.9%	10.2	94.6
Ours SARDE Inception	35.8%	58.7%	**6.1**	**31.5**
Ours SARDE Series	36.7	60.1%	**5.8**	**29.2**

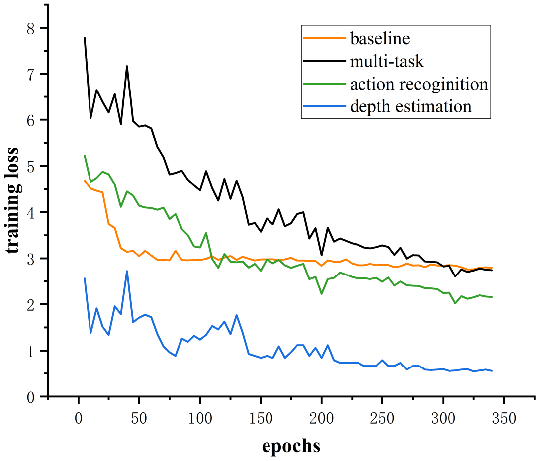


Fig. 7 The training loss with epochs, SARDE with baseline ST-GCN. The multi-task loss $\mathcal{L}$ comprises the depth estimation loss $\mathcal{L}_I$ and the action recognition loss $\mathcal{L}_{II}$ defined in Eq. (3); the baseline loss is equivalent to the action recognition loss defined in Eq. (6).

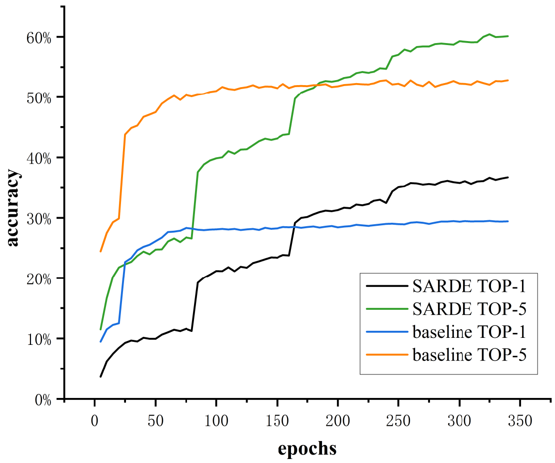


Fig. 8 The action recognition accuracy with epochs, SARDE with baseline ST-GCN. The accuracy TOP-1 and TOP-5 are defined in Eqs. (8) and (9).

first attempt at human skeleton action recognition integrated with depth estimation and has achieved good results. We are confident this will be an important research topic in related scenarios.

The visualization comparison results with the baseline ST-GCN [37] are shown in Figs. 7 and 8.

- The training error of the baseline model converges faster in the initial 120 epochs, in that in Fig. 7, the green curve is mainly located above the orange curve. The proper reason is that, compared to the baseline model, the multi-task SARDE method model is larger in size due to the addition of the depth estimation task, resulting in more learning parameters and training epochs.
- The training loss of the proposed SARDE continues to converge at a faster rate after epoch 50, as evidenced by the indication in Fig. 7 that after epoch 150, the green curve gradually falls below the orange curve, and the range between them tends to increase, showing that as the number of iterations increases, the multi-task training continues to drive the network model towards the optimized target.
- The proposed SARDE requires more iteration epochs to finish the training than the baseline ST-GCN. In Fig. 8, the accuracies of the ST-GCN reach practically flat after epoch 50, indicating that its training has been completed; the accuracies of the SARDE tend to level after epoch 300. This result confirms that with the increase in model complexity and size, it takes more epochs and time for network training to achieve its optimization.
- The proposed SARDE can achieve better HAR performance compared to the baseline ST-GCN. As in Fig. 7, the action recognition loss in green is below the baseline loss in black, and in Fig. 8, the accuracies of the SARDE have surpassed that of the ST-GCN.

5. Conclusion

A 2D human skeleton action recognition method based on depth estimation (SARDE) is proposed in this study to increase the accuracy of HAR in video surveillance with 2D information captured with monocular cameras by transforming 2D human action into 3D format. Depth estimation networks with series and inception blocks are developed to be integrated with action recognition networks for multi-task learning in an end-to-end manner to fully exploit the correlation between action recognition and depth estimation by sharing network parameters, such that the depth feature of the human skeleton joint can be more effectively learned to achieve human action recognition. The experimental results verify the effectiveness and superiority of the proposed model, and the model reaches state-of-the-art on the Kinetics dataset. This study has been the first attempt at HAR integrated with human skeleton depth estimation and has achieved good results. We are confident this will be an important research topic in the related scenarios. In future work, SARDE can serve as a vital method for intelligent visual surveillance and HRC robot control in multi-model fusion.

References

[1] S. Gu, L. Wang, L. He, X. He, and J. Wang, "Gaze estimation via a differential eyes' appearances network with a reference grid," Engineering, vol.7, no.6, pp.777–786, 2021.

[2] G. Saleem, U.I. Bajwa, and R.H. Raza, "Toward human activity recognition: A survey," Neural Computing and Applications, vol.35, no.5, pp.4145–4182, 2023.

[3] C.N. Phyo, T.T. Zin, and P. Tin, "Deep learning for recognizing human activities using motions of skeletal joints," IEEE Trans. Consum. Electron., vol.65, no.2, pp.243–252, 2019.

[4] P.F. Felzenszwalb and D.P. Huttenlocher, "Pictorial structures for object recognition," International Journal of Computer Vision, vol.61, no.1, pp.55–79, 2005.

[5] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," Advances in Neural Information Processing Systems, vol.27, 2014.

[6] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," 2016 Fourth International Conference on 3D Vision (3DV), pp.239–248, IEEE, 2016.

[7] C. Godard, O.M. Aodha, and G.J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.270–279, 2017.

[8] K. Wang, L. Lin, C. Jiang, C. Qian, and P. Wei, "3D human pose machines with self-supervised learning," IEEE Trans. Pattern Anal. Mach. Intell., vol.42, no.5, pp.1069–1082, 2019.

[9] X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional human pose regression," Proc. IEEE International Conference on Computer Vision, pp.2602–2611, 2017.

[10] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," arXiv preprint, arXiv:1511.07122, 2015.

[11] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," Proc. IEEE/CVF International Conference on Computer Vision, pp.12179–12188, 2021.

[12] L. Wang, J. Zhang, Y. Wang, H. Lu, and X. Ruan, "CLIFFNet for monocular depth estimation with hierarchical embedding loss," European Conference on Computer Vision, pp.316–331, Springer, 2020.

[13] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "DenseNet: Implementing efficient ConvNet descriptor pyramids," arXiv preprint, arXiv:1404.1869, 2014.

[14] H. Jung, Y. Kim, D. Min, C. Oh, and K. Sohn, "Depth prediction from a single image with conditional adversarial networks," 2017 IEEE International Conference on Image Processing (ICIP), pp.1717–1721, IEEE, 2017.

[15] J. Jiao, Y. Cao, Y. Song, and R. Lau, "Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss," Proc. European Conference on Computer Vision (ECCV), pp.55–71, 2018.

[16] C.N. Phyo, T.T. Zin, and P. Tin, "Skeleton motion history based human action recognition using deep learning," 2017 IEEE 6th Global Conference on Consumer Electronics (GCCE), pp.1–2, IEEE, 2017.

[17] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.3633–3642, 2017.

[18] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-based action recognition with hierarchical spatial reasoning and temporal stack learning network," Pattern Recognition, vol.107, 107511, 2020.

[19] Y. Liu, H. Zhang, Y. Li, K. He, and D. Xu, "Skeleton-based human action recognition via large-kernel attention graph convolutional network," IEEE Trans. Vis. Comput. Graph., vol.29, no.5, pp.2575–2585, 2023.

[20] J. Zhang, Z. Tu, J. Yang, Y. Chen, and J. Yuan, "MixSTE: Seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.13232–13242, 2022.

[21] J. Zhang, G. Ye, Z. Tu, Y. Qin, Q. Qin, J. Zhang, and J. Liu, "A spatial attentive and temporal dilated (SATD) GCN for skeleton-based action recognition," CAAI Trans. Intelligence Technology, vol.7, no.1, pp.46–55, 2021.

[22] S. Cho, M.H. Maqbool, F. Liu, and H. Foroosh, "Self-attention network for skeleton-based human action recognition," Proc. IEEE/CVF Winter Conference on Applications of Computer Vision, pp.635–644, 2020.

[23] Z. Tu, J. Zhang, H. Li, Y. Chen, and J. Yuan, "Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition," IEEE Trans. Multimed., vol.25, pp.1819–1831, 2022.

[24] B. Xu, X. Shu, and Y. Song, "X-invariant contrastive augmentation and representation learning for semi-supervised skeleton-based action recognition," IEEE Trans. Image Process., vol.31, pp.3852–3867, 2022.

[25] K. Su, X. Liu, and E. Shlizerman, "Predict & cluster: Unsupervised skeleton based action recognition," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.9631–9640, 2020.

[26] S. Chaudhary and S. Murala, "Depth-based end-to-end deep network for human action recognition," IET Computer Vision, vol.13, no.1, pp.15–22, 2019.

[27] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.7291–7299, 2017.

[28] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.3595–3603, 2019.

[29] T. Kipf, E. Fetaya, K.C. Wang, M. Welling, and R. Zemel, "Neural relational inference for interacting systems," International Conference on Machine Learning, pp.2688–2697, PMLR, 2018.

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K.Q. Weinberger, "Densely connected convolutional networks," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.2261–2269, 2017.

[31] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," Proc. AAAI Conference on Artificial Intelligence, vol.31, no.1, 2017.

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

[33] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A.C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," IEEE Trans. Pattern Anal. Mach. Intell., vol.42, no.10, pp.2684–2701, 2019.

[34] W. Yang, J. Zhang, J. Cai, and Z. Xu, "HybridNet: Integrating GCN and CNN for skeleton-based action recognition," Applied Intelligence, vol.53, no.1, pp.574–585, 2023.

[35] K. Hu, J. Jin, C. Shen, M. Xia, and L. Weng, "Attentional weighting strategy-based dynamic GCN for skeleton-based action recognition," Multimedia Systems, vol.29, pp.1941–1954, 2023.

[36] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," Proc. European Conference on Computer Vision (ECCV), pp.365–381, 2018.

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

[38] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.12026–12035, 2019.

[39] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.7912–7921, 2019.

[40] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," IEEE Trans. Image Process., vol.29, pp.9532–9545, 2020.

[41] N. Sun, L. Leng, J. Liu, and G. Han, “Multi-stream slowfast graph convolutional networks for skeleton-based action recognition,” Image and Vision Computing, vol.109, 104141, 2021.

[42] J. Xie, Q. Miao, R. Liu, W. Xin, L. Tang, S. Zhong, and X. Gao, “Attention adjacency matrix based graph convolutional networks for skeleton-based action recognition,” Neurocomputing, vol.440, pp.230–239, 2021.

[43] L. Wu, C. Zhang, and Y. Zou, “Spatiotemporal focus for skeleton-based action recognition,” Pattern Recognition, vol.136, 109231, 2023.

Lei Wang received her M.S. degree in Pattern Recognition and Intelligence Systems from Beihang University, Beijing, China, in 2010. She is currently a Ph.D. student at Sichuan University, Chengdu, China. Her major is Information and Communication Engineering.

Shanmin Yang received her Ph.D. degree from Sichuan University, Chengdu, China, in 2021. She is currently a lecturer at the School of Computer Science, Chengdu University of Information Technology, Chengdu, China. Her research interests include intelligent meteorology, machine learning, and computer vision.

Jianwei Zhang received Bachelor, M.S. and Ph.D. degrees in Computer Technology from Sichuan University, Chengdu, China, in 1993, 2000, and 2008. He is currently a professor and doctoral supervisor at Sichuan University. He is a member of the Expert Committee of the China Virtual Reality and Visualization Technology Alliance.

Song Gu received the M.S. and Ph.D. degrees in engineering from the University of Electronic Science and Technology of China in 2016. He has been a Professor at Chengdu Aeronautic Polytechnic since 2022. He researched with the KTH Royal Institute of Technology as a Visiting Scholar in 2019. His main research interest is video object tracking and segmentation.