Computing and Artificial Intelligence 2025, 3(3), 2342.
doi: https://doi.org/10.59400/cai2342

Article

Joint weight adversarial attack based on human skeleton action recognition

Haopeng Mu, Ling Guo*, Xiaozhou Zhang

School of Information Science and Technology, Northwest University, Xi'an 710500, China
*** Corresponding author:** Ling Guo, lingo@nwu.edu.cn

CITATION

Mu H, Guo L, Zhang X. Joint weight adversarial attack based on human skeleton action recognition. Computing and Artificial Intelligence. 2025; 3(3): 2342. https://doi.org/10.59400/cai2342

ARTICLE INFO

Received: 23 December 2024
Revised: 16 May 2025
Accepted: 22 May 2025
Available online: 4 July 2025

Abstract: In recent years, due to the excellent spatial information correlation, small data volume and high computational efficiency of bone data, it has been widely applied in action recognition fields such as autonomous driving and intelligent security. However, in practical applications, attackers only need to apply a small perturbation to the input bone data to cause the attacked model to make incorrect recognition of the corresponding action, thereby resulting in a significant drop in recognition accuracy and even potentially causing serious consequences in high-risk scenarios such as autonomous driving. To solve this problem, many attack methods have been proposed, such as attacks that limit the angle changes between bones or attacks that alter the length of bones. These methods can, to a certain extent, increase the attack success rate of action recognition models, but most of these methods attack the bone data by simply disregarding the influence of each joint bone node on the overall action. In this paper, we propose a new adversarial attack method, that is, to attack through interfering with the coordinate data of the entire skeletal joint nodes. In our method, the concept of joint weights is proposed, and a time cropping translation attack is designed based on joint weights to improve the attack success rate. We conducted experiments on our method. The experimental results show that our attack success rate is stable at over 60%.

Keywords: human action recognition; skeleton sequence; black-box attack; adversarial machine learning

1. Introduction

In recent years, due to the facts that bone data is not sensitive to individual characteristics and bone data can avoid the influence of external interference such as clothing occlusion, lighting influence, action blur and shooting background in the original RGB image or video, bone data is widely used in the field of action recognition [1–4]. However, bone data has less redundancy such that it is extremely sensitive to disturbances. The continuity of actions in the time dimension is very strong and there is a strong correlation between its adjacent nodes. Besides, intra-frame action information is also very rich. All these reasons make bone data extremely vulnerable to adversarial attacks such that the use of adversarial attacks in practical applications poses a huge challenge for the security of action recognition models based on bone data.

Current adversarial attack methods are mainly divided into two categories: white-box attacks and black-box attacks. A white box attack assumes that the attacker fully understands the structure and parameters of the target system and can design the corresponding attack mode according to the specific structure and parameters of the target system and carry out detailed analysis and optimization of the target system. The

realization conditions of this situation are very harsh and difficult to achieve in real life [5, 6]. Black box attacks can only obtain the input and output information of the system and attack the target system by modifying the input information, which can be achieved by using methods such as black box optimization. Most of the existing attack methods adopt black box attacks [7, 8]. These adversarial attack methods mainly carry out attacks by perturbing data nodes and controlling the disturbance in an imperceptible range by constraints [9, 10]. The main flow chart of the adversarial attack based on action recognition of bone data is shown in **Figure 1**. First, adversarial samples are input into the action recognition model to obtain a confidence vector. Then predicted labels are obtained from the confidence vectors. These predicted labels are compared with the original labels. If the results are inconsistent, then this attack is successful.

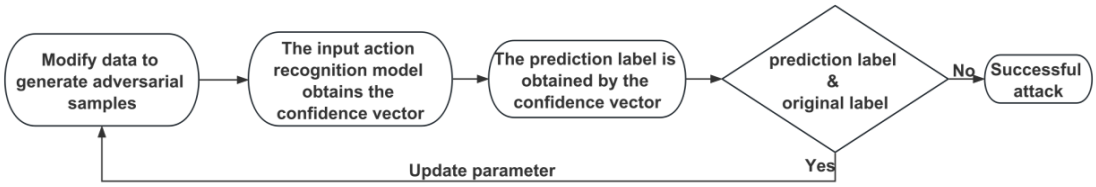

Figure 1. The main flow chart of adversarial attacks based on action recognition from bone data.

Note: The adversarial sample is obtained by perturbation and constraint, and it is input into the model to obtain the predicted label. If the predicted label is inconsistent with the original label, the attack is successful.

When modifying the original bone data to generate adversarial samples, the root node of the bone data must first be determined, which is the starting node of the attack. Since the underlying graph of human bone data is a tree, different root nodes will change the coordinates of the nodes connected to it [11–13]. After subsequent nodes are attacked, the impact on the attack result is different.

In addition, when the human body does different movements, each joint of the human body will make corresponding adjustments to match the overall movement change, which leads to the different roles of each joint node in a group of movements. Some joints may be the main body of the movement, and some may be the coordination and cooperation of the movement. In other words, each joint has a different degree of influence on the overall movement in a group of movements, which makes the same attack effect different on different joint nodes [14–16]. Therefore, in order to improve the effectiveness of the attack, the influence of each joint node on the action should be determined before the antagonistic attack. However, existing attack methods usually use the initial node specified in the data set as the root node to attack [17, 18]. If the root node is determined in this way, the attack effect of some actions may not be ideal or may even reduce the attack effect. Therefore, in order to maximize the attack effect, the choice of root node is particularly important. Moreover, these methods do not consider what role a joint node plays in the action, but rather what effect an attack will have on a joint node or on the action itself. All of these measures reduce the success rate of attacks. In this paper, the weight of a joint is determined by the influence of each node in a group of actions on the whole movement under the same conditions, and the data is spliced and reorganized according to the time dimension, so as to attack the data. In addition, the joint in a group of data with relatively small influence on the movement

is found as the root node through the joint weight. Therefore, a new adversarial attack method is proposed.

The main contributions in this paper are as follows:

(1) We propose the concept of joint weights. The joint weights are determined by comparing the impact of the change of each joint on the overall action. By using the joint weights, the attack strength can be adjusted appropriately for each joint during the attack.
(2) We propose a data clipping method matching joint weights, which can make the perturbed data more natural.
(3) We propose a root node numbering method to change the root nodes.

The rest of this article is structured as follows: We briefly review the related work in Section 2. In Section 3, our attack method and implementation details are presented, and corresponding formulas are given. In Section 4, we give the experimental setup and experimental results and analyze the experimental results. We give our conclusions in Section 5.

2. Related work

2.1. Skeleton-based action recognition

Nowadays, action recognition technology has a wide range of applications in the fields of unmanned driving, intelligent monitoring and human-computer interaction. At present, action recognition technology based on bone data has been widely used in action recognition [19], and has gradually become an important means of action recognition technology.

There are three kinds of action recognition methods based on bone data: RNN-based, CNN-based and GCN-based.

The RNN-based approach represents the bone movement data as a time series of actions, each consisting of the coordinates of all the joints [20–22]. RNN has many advantages in action recognition based on bone data. For example, RNN can process sequence data, which is suitable for the modeling of time series such as bone data and can capture the temporal relationship in action sequences. However, it is difficult for RNNs to make use of the interaction of bone joints in space and time at the same time, and there are problems of gradient disappearance or gradient explosion.

The CNN-based method recognizes action data by using image classification, that is, converting bone movement data into a pseudo-image or joint trajectory map, represented as a two-dimensional tensor, where one dimension represents time and the other dimension superimposes all joints of a single bone [23–26]. CNN can effectively extract local features from images or bone data and has a good perception of local patterns of posture. It has good feature-level extraction ability. However, since skeleton data forms a graph structure in non-Euclidean space, with joints as vertices and natural body connections as edges, a lot of computing resources are wasted. Moreover, it is difficult for CNN to model time series data, and it is impossible to directly process the time sequence relationship of actions.

The approach based on GCN is to represent the skeleton topology, that is, joint

connectivity, in the form of a graph, where the nodes of the graph correspond to joints in the skeleton and the edges correspond to bones connecting adjacent joints, and graph convolution operations are applied to to identify actions [11–13, 27–30]. GCN can effectively extract information from bone data. For example, the spatial-temporal graph convolutional network (ST-GCN) [31] defines the GCN of bone data in terms of space. ST-GCN models bone joints and their natural connections within the same frame and takes joint connections between continuous moments as time edges. However, the adjacency matrix of ST-GCN is fixed; that is, it does not create connections that do not exist, and it does not take full advantage of the important features of bone length and orientation in the bone data.

2.2. Adversarial attacks on skeletons

Although the performance of action recognition has been greatly successful, and the recognition success rate and accuracy have also been greatly improved, it is undeniable that deep learning is vulnerable to well-designed adversarial attacks [6,8,32]. Perturbations in the data that are not obvious to humans can easily fool deep learning models [17], leading to significant security issues. And deep learning-based action recognition is also proving to be highly susceptible to adversarial perturbations in 2020, and since then, various adversarial attack methods based on action recognition based on bone data have been proposed. CIASA was the first one to propose an adversarial attack on action recognition models using bone data [33]. While limiting the size of joint disturbance and joint acceleration, the generated adversarial network is used to physically constrain joints and so on to carry out the disturbance attack. However, this method only disturbs the coordinates of joints and then forms a complete skeleton through constraints such as bone length and so on. Because the ST-GCN model does not take full advantage of the bone length feature, Zheng [34] proposes an attack that combines bone length, joint angle, and joints acceleration to define the enhanced Lagrange formula. Goodfellow et al. [18] proposed an adversarial attack by modifying the length of the skeleton, but only in the lower dimensions (around 30), and did not attack in the time dimension.

Therefore, this paper proposes a new attack method based on the above. First, we propose a concept of joint weight, which determines the impact of the current joint on the overall movement by comparing the ways of changing the joint value, and cuts the data by joint weight, so as to carry out the attack in the time dimension. At the same time, we repositioned the root node of the bone by joint weights and judged the stability of the joint by changing joint weights and coordinates, so as to re-determine the root node of the bone, instead of directly using the initial node specified by the data set as the root node as the existing method.

3. Method

In order to determine the priority of joint processing in a set of bone data, we put forward the concept of joint weight. After determining the priority of joint processing, the position of the root node is determined through this priority, and it is used as the basis for editing data. Our experimental principle is shown in **Figure 2**.

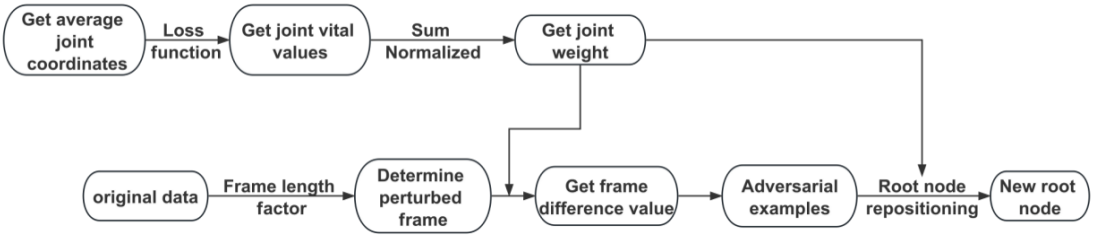

Figure 2. Study the flow chart of ideas.

Note: First, the average coordinate of each joint is used to replace the original coordinate, and the joint weight is obtained by the change degree of the loss function. Then, the joint weight is used to compare the determined attack frame with the attacked data to obtain the frame difference value, and the frame with the smallest frame difference value is connected to the attack frame to obtain the countermeasure sample. By combining the joint change degree and joint weight, the joint node with the least joint change degree is identified as the root node.

3.1. Preface

We assume that a bone dataset action recognition result has class L, and an action (skeleton sequence) can be expressed as $X = \{q_i^\tau \mid i = 1, 2, ..., N, \tau = 0, 1, ..., T-1\}$, where N, T are the number of joints and the number of frames of this action respectively, $q_i^\tau \in \mathbb{R}^3$ represents the three-dimensional coordinates of the i-th joint of the τ-th frame of the bone data in this action. After action X passes through the attacked model f, an L-dimensional confidence vector $f(X)_k$ will be generated, representing the confidence score of this action in the K-class action. The class label of an action in the form of one-hot can be represented as $y = (y_1, y_2, ..., y_L)^T$.

3.2. Joint weight

We first average the coordinates of all frames of the *i* joint q_i in a set of data to obtain the average position M_i of the *i* joint in this set of data, as shown in Equation (1):

$$M_i = \frac{\sum_{\tau=0}^{N} q_i^\tau}{N} \tag{1}$$

After the average position M_i is obtained, the coordinates of the *i*-th joint and the average position of each frame are averaged again to obtain the new data $\widetilde{X}$. The difference between the loss value obtained by sending the new data into the model and the loss value of the original data X is the important value im_i of the *i*-th joint, as shown in Equation (2):

$$im_i = L\left(f\left(\frac{M_i + x_i}{2}\right)\right) - L\left(f\left(x_i\right)\right) \tag{2}$$

where L is a loss function, the larger the im_i value is, the greater the impact on this set of data after the change of the *i*-th joint, and the more important it is to this set of data.

Then we sum and normalize all the joint important values of this set of data to get the proportion of each joint important value in the whole, which is the joint weight. The corresponding formula is shown in Equation (3):

$$W_i = \frac{im_i}{\sum_{i=0}^{J} im_i} \tag{3}$$

3.3. Time clipping translation attack method

In order to realize the time clipping translation attack method, we need to use the joint weights mentioned above, through which we can focus on the comparison of each joint, the implementation method is as follows: First, we determine the length M of the attack segment in F frame by Equation (4):

$$M = F \times \theta, \quad \theta \in N(0, \epsilon) \tag{4}$$

where θ represents a random value sampled from the normal distribution (0, ϵ). It is one of the parameters controlling the disturbance size in the PGD algorithm, which means that the difference between the pixel value of each point of the admixture sample and the original image does not exceed ϵ by using the projection function. When the value of M is too large, the change of the action after the attack is unnatural and the imperceptibility of the attack is reduced.

The second step is to randomly determine an attack start frame fr^{start} from the remaining frames in the original data, except for the attack frame. That is to say, starting from the attack start frame fr^{start}, the next attack segment of length M is cropped to obtain the attack end frame fresh. In the third step, we compare the start and end positions of the attack segment, namely the attack start frame fr^{start} and the attack end frame fr^{end}, with all adjacent frames of the remaining N-M frames. We find the adjacent frames f^{τ} and $fr^{\tau+1}$ with the smallest difference value, and then connect fr^{τ} to fr^{start}, $fr^{\tau+1}$ to fr^{end}, and the original $fr^{start-1}$ to fr^{end+1} to obtain a new sample X'. The corresponding formula is shown in Equation (5):

$$diff^{\tau} = \sum_{i=0}^{J} \left(W_i \cdot sqrt\left(\left(fr_i^{\tau} - fr_i^{start} \right)^2 + \left(fr_i^{\tau+1} - fr_i^{end} \right)^2 \right) \right) \tag{5}$$

where $diff^{\tau}$ represents the difference between frame τ and attack start frame fr^{start}, frame $\tau + 1$ and attack end frame fr^{end}. By comparing the difference between the attack start frame fr^{start} and the end frame fr^{end} of each joint, and the frame fr^{τ} to be compared and the frame $fr^{\tau+1}$ of the next frame, we can get the transformation between the corresponding two sets of frames, and then multiply the weight of the joint to get the difference value $diff^{\tau}$ between the attack segment and the compared segment. The smaller the value of $diff^{\tau}$ is, the smaller the difference between frames τ and $\tau + 1$ and the attack segment. The frame with the smallest difference value is taken as the connection frame.

3.4. Root node relocation method

After the time clipping translation attack, we repositioned the root node of each set of data. The existing methods did not redetermine the position of the root node when disturbing the bone data but used the initial node of the data set itself as the root node, while most of the existing data sets usually used the spine root node or the chest joint node as the root node. However, if the attack mode is to traverse the nodes successively from the root node, then the fixed node may not improve the success rate of the attack

or even reduce the success rate of the attack and may reduce the imperceptibility of the attack, because the fixed root node may change greatly in some movements, resulting in a large gap between the intra-frame joint coordinates and the original coordinates after the attack and poor inter-frame continuity. The attack success rate decreases. In order to solve this problem, we propose a root node relocation method, which combines the coordinate change amplitude of the joint with the joint weight as the basis for judging whether the root node can be used as the root node. The specific method is shown in Equation (6):

$$dis_i = w_i \cdot \left(\sum_{\tau=0}^{N} \sqrt{\left(x_i^{\tau} - x_i^{\tau-1}\right)^2 + \left(y_i^{\tau} - y_i^{\tau-1}\right)^2 + \left(z_i^{\tau} - z_i^{\tau-1}\right)^2} \right) \tag{6}$$

where dis_i indicates the degree of position change of the i-th joint in this movement, and the smaller the dis_i value, the more stable the joint is and the more suitable it is to be the root node.

Using the root node relocation method to attack data will change the root node position of the original action, which will also change the attack order of the joint and will have different impacts on the subsequent joints of each joint. Using the time clipping translation attack method will change the impact on the action in the time dimension, resulting in corresponding changes in the continuity of the action, thus changing the recognition result. The specific effect is shown in **Figure 3**.

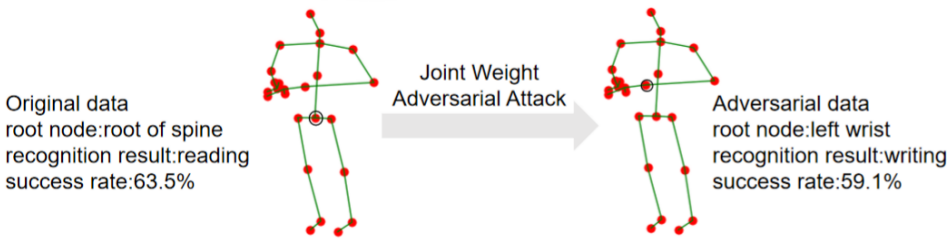

Figure 3. After the joint weight adversarial attack, the original data recognition result changed from 63.5% reading to 59.1% writing, and the root node changed from the root of the spine to the left wrist.

4. Experiments and results

We first describe our experimental setup and evaluation metrics in sub-section 4.1, and then give the experimental results of our proposed attack in sub-section 4.2.

4.1. Experimental setup

4.1.1. Data set and test model

Our experiment used the NTU RGB+D dataset [35] and the ST-GCN model. The NTU RGB+D dataset is a three-dimensional skeleton action dataset, which divides human skeletons into 25 nodes and indexes each node. The human skeleton can be determined by the index number of the node connected to it. The data set was composed of 40 people aged 10 to 35 who were collected by three cameras in different locations, and the data set was divided into two criteria: Cross-Subject and Cross-View.

Cross-Subject divides the training set and the test set by personnel numbers, in which 20 people whose personnel numbers are 1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, and 38 are taken as the training set. With 40,320 samples, the rest of the personnel numbers serve as a test set, with 16,560 samples. Cross-View divides the training set and the test set by camera number. The samples collected by camera 1 are used as the test set, and cameras 2 and 3 are used as the training set. The sample numbers are 18,960 and 37,920 respectively.

4.1.2. Evaluation metrics and others

Next, we use our attack method to attack the test model and then evaluate the attack effect according to the attack success rate, and use the confidence score and the denormalized residual to compare the attack efficiency. Our experiments were conducted using an Intel Core i7-11700 CPU and an NVIDIA RTX 3060 GPU.

4.2. Experimental results

In order to verify the effectiveness of the attack, we conducted an attack experiment on the model, and the experimental results are shown in **Figure 4**.

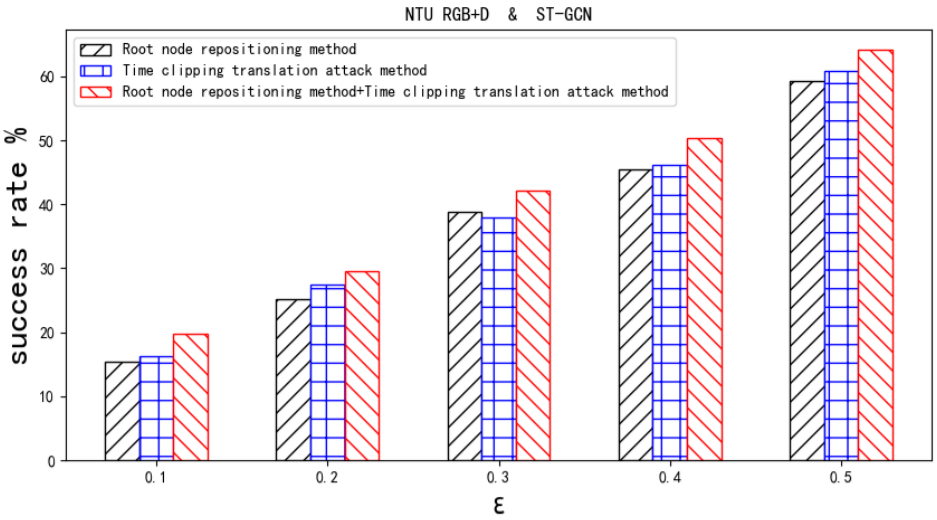

Figure 4. The result of attack experiment on the model.

Note: The effectiveness of the attack method against the ST-GCN model on the NTU RGB+D data set is evaluated by comparing the attack success rate.

When the value is $\varepsilon = 0.1$ and $\varepsilon = 0.2$, the attack success rate exceeds 15% and 25% respectively, because the disturbance to the data caused by the value is too small, and the impact on the identification result is very limited. However, in this case, our attack rate can still achieve a medium attack success rate, which proves that our attack is very effective in the range of small disturbance. When $\varepsilon = 0.5$, the attack success rate even exceeds 60%. Through comparison, we find that the success rate of the attack using the root node relocation method alone is close to that of the time clipping translation attack, but the success rate of the root node relocation method alone is slightly lower, which proves to some extent that the effect of the initial node change caused by the root node relocation method is relatively worse. The results of using the two methods at the same time will be greatly improved, which also shows that the combination of the two methods will greatly improve the efficiency of the attack.

In most cases, the changes of human movement are interrelated, and the change amplitude of the main part of the movement increases from the root node to the leaf node in turn, while the change degree of the non-main part of the movement also accords with the increase in the change degree from the root node to the leaf node in turn, but the increase degree is similar, and there is no big difference, resulting in similar joint weights. In the case that the main part of some actions contains fewer nodes (such as jumping in place, reading, writing, etc.), the overall joint weight differentiation is not large, which makes the role of the root node not very obvious, and the effect of the root node relocation method is relatively poor. The time clipping attack is to change the order of actions in the time dimension, and the joint weight is only the basis for judging the correlation of actions, and even the worse the correlation, the greater the degree of movement change, the higher the success rate of the attack, but this will lead to the reduction of imperceptibility. In order to ensure the imperceptibility of the attack, it is necessary to ensure the correlation between the attack segment and the attacked segment is as high as possible, which limits the success rate of the attack.

Next, we conducted a further experiment to analyze the effectiveness of the attack by comparing the change of confidence score. The experimental results are shown in **Figure 5**.

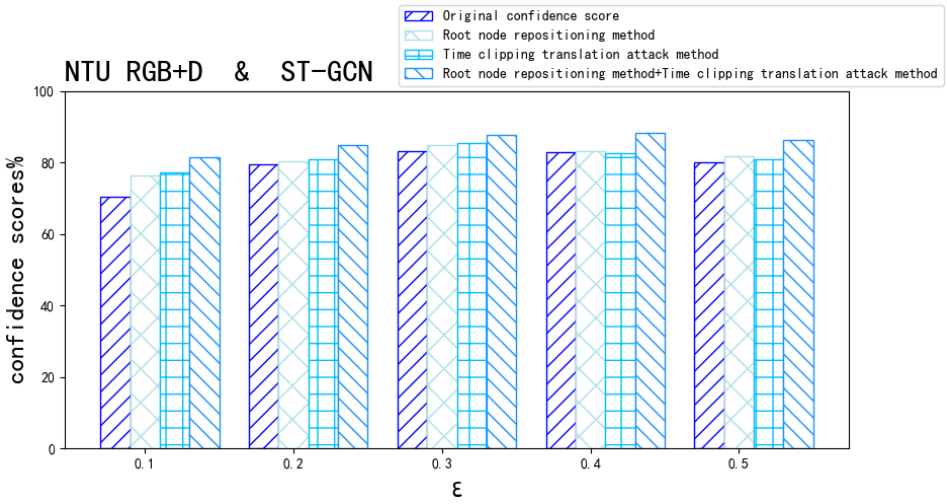

Figure 5. The result of comparing the change of confidence score.
Note: We use confidence scores to compare the effectiveness of our methods. Using the two methods alone is effective, but the improvement of the confidence score is not great, while using the two methods at the same time has a large degree of improvement in the confidence score.

Through the comparison of confidence scores, we can see that the confidence scores of the root node relocation method and the time clipping translation attack are similar to and slightly higher than the original confidence score, while the confidence scores of the two methods are significantly higher than the original confidence score, which also proves the effectiveness of our attack method from the side. However, the confidence scores of the root node relocation method and the time clipping translation attack are very close, so we conducted a more detailed comparison experiment on the confidence scores of the denormalized residuals, and the experimental results are shown in **Figure 6**.

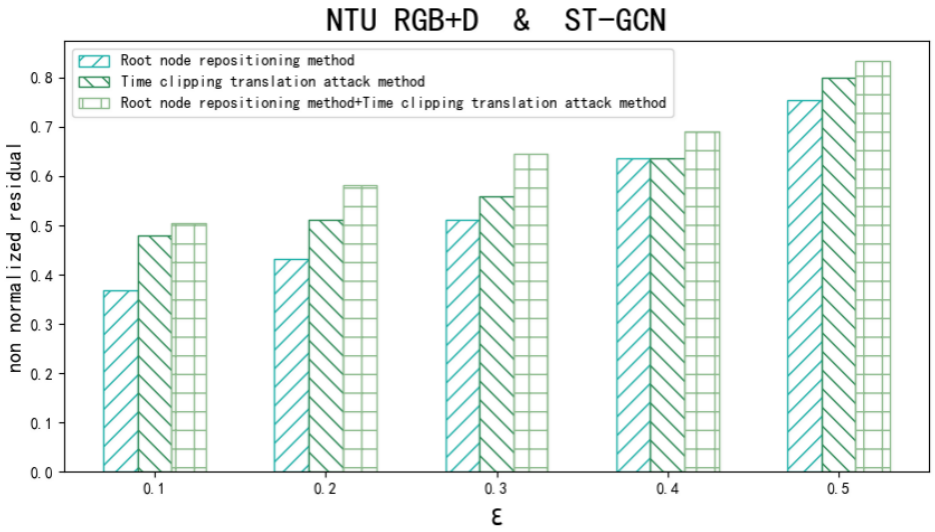

Figure 6. The experiment results of non-normalized residuals.

Note: It can be seen that the experimental result of using root node relocation alone is relatively poor. The experimental results of using time clipping translation attacks alone are better, which also proves that the effectiveness of the attack is better. At the same time, using these two methods can significantly change the confidence value, so that the confidence value after the attack changes greatly, indicating that the attack is effective and stable.

The denormalized residual is an indicator to measure the gap between the confidence value after the attack and the original confidence value. The specific calculation equation is shown in Equation (7):

$$nq = \frac{max_{x'}\{p(f(x') \mid x')\} - p(y \mid x)}{\max_{x'}\{p(f(x') \mid x')\}} \tag{7}$$

where f is the attacked model, x is the original sample, x' is the admissible sample, y is the original label, $p(f(x')|x')$ is the confidence value of the admissible sample, and $max_{x'}\{p(f(x')|x')\}$ is the maximum confidence value of the admissible sample. The denormalized residuals pay more attention to the change of confidence value after the attack. The larger the value is, the more effective the adversarial attack is to the model, and the less robust the model is to the adversarial sample.

As can be seen from **Figure 5**, the experimental results of root node relocation alone are relatively poor, which also proves that only changing the initial position of the attack node has a relatively small impact on the experimental results, while the experimental results of time clipping translation attack alone will be improved to a large extent, which also proves the effectiveness of time clipping translation attack. The main reasons for the difference between the root node relocation method alone and the time-clipping translation attack are similar to the analysis in **Figure 3**. The experimental results of using both methods are even better, which also proves that our attack method is successful.

We then compared the existing adversarial attack methods on the NTU RGB+D dataset, and the comparison results are shown in **Table 1**.

As can be seen from **Table 1**, if only the root node relocation method is used, the attack success rate is similar to the results of existing methods and has a certain improvement compared with some methods, but the improvement degree is not high. The success rate of using time clipping translation attacks has been greatly improved, and it can be said that it has exceeded the success rate of most existing methods. The

performance of the joint weight adversarial attack is the best, reaching 66.1%, which proves the effectiveness of our method, indicating that the choice of root node is very important and the attack on the data from the time dimension is effective.

Table 1. The comparison of current adversarial attack methods based on human bone data with ours using the NTU RGB+D dataset.

Attack method	NTU RGB+D success rate
2s-AGCN	57.9%
P3D	61.8%
ABLAAR	59.8%
CIASA	58.4%
BASAR	58.13%
Root node relocation method	59.2%
Time clipping translation attack method	60.9%
JWAA	66.1%

5. Conclusions

In this paper, a concept of joint weight is proposed, which focuses on the disturbance of joints by comparing the importance of human bone data nodes to the whole movement. Based on the joint weight, an attack method of cutting and splicing bone data in the time dimension is designed, and the effectiveness of the attack is proved. At the same time, in order to solve the problem that the existing attack methods simply determine the initial node specified in the data set as the root node without proper modification according to the specific attack mode, which may lead to poor attack efficiency and stability, we designed a root node relocation method according to the joint weight to solve the problem that the root node does not adapt to the attack mode.

Author contributions: Conceptualization, HM and LG; methodology, HM; software, HM; validation, HM, XZ and LG; formal analysis, HM and XZ; investigation, HM and XZ; resources, HM and XZ; data curation, HM and XZ; writing—original draft preparation, HM; writing—review and editing, HM and LG; visualization, HM; supervision, HM and LG; project administration, HM. All authors have read and agreed to the published version of the manuscript.

Funding: This work received no external funding.

Institutional review board statement: Not applicable.

Informed consent statement: Not applicable.

Data availability statement: No new data were created.

Conflict of interest: The authors declare no conflict of interest.

References

1. Chen G, Chenb S, Fan L, et al. Who is Real Bob? Adversarial Attacks on Speaker Recognition Systems. In: Proceedings of the 2021 IEEE Symposium on Security and Privacy (SP); 24–27 May 2021; San Francisco, CA, USA. pp. 694–711. doi: 10.1109/SP40001.2021.00004

2. Chen Z, Xie L, Pang S, et al. Appending Adversarial Frames for Universal Video Attack. In: Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV); 3 January 2021; Waikoloa, HI, USA. pp. 3198–3207. doi: 10.1109/WACV48630.2021.00324
3. Tanaka N, Kera H, Kawamoto K. Adversarial Bone Length Attack on Action Recognition. Proceedings of the AAAI Conference on Artificial Intelligence. 2022; 36(2): 2335–2343. doi: 10.1609/aaai.v36i2.20132
4. Kong J, Deng H, Jiang M. Symmetrical Enhanced Fusion Network for Skeleton-Based Action Recognition. IEEE Transactions on Circuits and Systems for Video Technology. 2021; 31(11): 4394–4408. doi: 10.1109/TCSVT.2021.3050807
5. Goodfellow IJ, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples. arXiv preprint. 2014. doi: 10.48550/ARXIV.1412.6572
6. Gowal S, Qin C, Uesato J, et al. Uncovering the Limits of Adversarial Training against Norm-Bounded Adversarial Examples. arXiv preprint. 2020. doi: 10.48550/ARXIV.2010.03593
7. Cheng K, Zhang Y, He X, et al. Extremely Lightweight Skeleton-Based Action Recognition With ShiftGCN++. IEEE Transactions on Image Processing. 2021; 30: 7333–7348. doi: 10.1109/TIP.2021.3104182
8. Diao Y, Shao T, Yang Y-L, et al. BASAR:Black-box Attack on Skeletal Action Recognition. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 20 June 2021; Nashville, TN, USA. pp. 7593–7603. doi: 10.1109/CVPR46437.2021.00751
9. Liu J, Akhtar N, Mian A. Adversarial Attack on Skeleton-Based Human Action Recognition. IEEE Transactions on Neural Networks and Learning Systems. 2022; 33(4): 1609–1622. doi: 10.1109/TNNLS.2020.3043002
10. Pony R, Naeh I, Mannor S. Over-the-Air Adversarial Flickering Attacks against Video Recognition Networks. arXiv preprint. 2020. doi: 10.48550/ARXIV.2002.05123
11. Zhang X, Xu C, Tao D. Context Aware Graph Convolution for Skeleton-Based Action Recognition. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 29 June 2020; Seattle, WA, USA. pp. 14321–14330. doi: 10.1109/CVPR42600.2020.01434
12. Li M, Chen S, Chen X, et al. Actional-Structural Graph Convolutional Networks for Skeleton-Based Action Recognition. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 15–20 June 2019; Long Beach, CA, USA. pp. 3590–3598. doi: 10.1109/CVPR.2019.00371
13. Shi L, Zhang Y, Cheng J, et al. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 15–20 June 2019; Long Beach, CA, USA. pp. 12018–12027. doi: 10.1109/CVPR.2019.01230
14. Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 18 June 2018; Salt Lake City, UT, USA. pp. 7132–7141. doi: 10.1109/CVPR.2018.00745
15. Wu H, Liu J, Zha Z-J, et al. Mutually Reinforced Spatio-Temporal Convolutional Tube for Human Action Recognition. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence; 10 August 2019; Macao, China. pp. 968–974. doi: 10.24963/ijcai.2019/136
16. Cho S, Maqbool MH, Liu F, et al. Self-Attention Network for Skeleton-based Human Action Recognition. In: Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV); 1 March 2020; Snowmass Village, CO, USA. pp. 624–633. doi: 10.1109/WACV45572.2020.9093639
17. Fursov I, Zaytsev A, Burnyshev P, et al. A Differentiable Language Model Adversarial Attack on Text Classifiers. IEEE Access. 2022; 10: 17966–17976. doi: 10.1109/ACCESS.2022.3148413
18. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Communications of the ACM. 2020; 63(11): 139–144. doi: 10.1145/3422622
19. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv preprint. 2014. doi: 10.48550/ARXIV.1412.6980
20. Du Y, Wang W, Wang L. Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 17 June 2015; Boston, MA, USA. pp. 1110–1118. doi: 10.1109/CVPR.2015.7298714
21. Liu J, Shahroudy A, Xu D, et al. Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition. In: Leibe B, Matas J, Sebe N, et al. (editors). Computer Vision–ECCV 2016, Lecture Notes in Computer Science. Springer International Publishing; 2016. pp. 816–833. doi: 10.1007/978-3-319-46487-9_50
22. Song S, Lan C, Xing J, et al. An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data. Proceedings of the AAAI Conference on Artificial Intelligence. 2017; 31(1). doi: 10.1609/aaai.v31i1.

11212

23. Zhang P, Lan C, Xing J, et al. View Adaptive Neural Networks for High Performance Skeleton-Based Human Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019; 41(8): 1963–1978. doi: 10.1109/TPAMI.2019.2896631
24. Ke Q, Bennamoun M, An S, et al. A New Representation of Skeleton Sequences for 3D Action Recognition. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 7 July 2017; Honolulu, HI, USA. pp. 4570–4579. doi: 10.1109/CVPR.2017.486
25. Liu M, Liu H, Chen C. Enhanced skeleton visualization for view invariant human action recognition. Pattern Recognition. 2017; 68: 346–362. doi: 10.1016/j.patcog.2017.02.030
26. Kim TS, Reiter A. Interpretable 3D Human Action Analysis with Temporal Convolutional Networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 21 July 2017; Honolulu, HI, USA. pp. 1623–1631. doi: 10.1109/CVPRW.2017.207
27. Cheng K, Zhang Y, He X, et al. Skeleton-Based Action Recognition With Shift Graph Convolutional Network. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 13–19 June 2020; Seattle, WA, USA. pp. 180–189. doi: 10.1109/CVPR42600.2020.00026
28. Liu Z, Zhang H, Chen Z, et al. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 13–19 June 2020; Seattle, WA, USA. pp. 140–149. doi: 10.1109/CVPR42600.2020.00022
29. Shi L, Zhang Y, Cheng J, et al. Skeleton-Based Action Recognition With Directed Graph Neural Networks. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 15 June 2019; Long Beach, CA, USA. pp. 7904–7913. doi: 10.1109/CVPR.2019.00810
30. Zhang P, Lan C, Zeng W, et al. Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 13–19 June 2020; Seattle, WA, USA. pp. 1109–1118. doi: 10.1109/CVPR42600.2020.00119
31. Yan S, Xiong Y, Lin D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. Proceedings of the AAAI Conference on Artificial Intelligence. 2018; 32(1). doi: 10.1609/aaai.v32i1.12328
32. Carlini N, Wagner D. Towards Evaluating the Robustness of Neural Networks. In: Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP); 28 May 2017; San Jose, CA, USA. pp. 39–57. doi: 10.1109/SP.2017.49
33. Krizhevsky A. Learning Multiple Layers of Features from Tiny Images. University of Toronto; 2009.
34. Zheng T, Liu S, Chen C, et al. Towards Understanding the Adversarial Vulnerability of Skeleton-based Action Recognition. arXiv preprint. 2020. doi: 10.48550/ARXIV.2005.07151
35. Shahroudy A, Liu J, Ng T-T, et al. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 26 June 2016; Las Vegas, NV, USA. pp. 1010–1019. doi: 10.1109/CVPR.2016.115