Adversarial Bone Length Attack on Action Recognition

Nariki Tanaka,[1] Hiroshi Kera,[2] Kazuhiko Kawamoto,[2]

[1] Graduate School of Science and Engineering, Chiba University
[2] Graduate School of Engineering, Chiba University
{ntanaka, kera}@chiba-u.jp, kawa@faculty.chiba-u.jp

Abstract

Skeleton-based action recognition models have recently been shown to be vulnerable to adversarial attacks. Compared to adversarial attacks on images, perturbations to skeletons are typically bounded to a lower dimension of approximately 100 per frame. This lower-dimensional setting makes it more difficult to generate imperceptible perturbations. Existing attacks resolve this by exploiting the temporal structure of the skeleton motion so that the perturbation dimension increases to thousands. In this paper, we show that adversarial attacks can be performed on skeleton-based action recognition models, even in a significantly low-dimensional setting without any temporal manipulation. Specifically, we restrict the perturbations to the lengths of the skeleton's bones, which allows an adversary to manipulate only approximately 30 effective dimensions. We conducted experiments on the NTU RGB+D and HDM05 datasets and demonstrate that the proposed attack successfully deceived models with sometimes greater than 90% success rate by small perturbations. Furthermore, we discovered an interesting phenomenon: in our low-dimensional setting, the adversarial training with the bone length attack shares a similar property with data augmentation, and it not only improves the adversarial robustness but also improves the classification accuracy on the original data. This is an interesting counterexample of the trade-off between adversarial robustness and clean accuracy, which has been widely observed in studies on adversarial training in the high-dimensional regime.

Introduction

Deep neural network models are highly vulnerable to adversarial perturbations, which are small input perturbations intentionally applied by an attacker (Szegedy et al. 2014). This poses a security concern in the use of deep neural network models in practical scenarios. Szegedy et al. (2014) was the first to discover the adversarial attack, that is, applying small perturbations to images that are imperceptible to a human but can fool deep neural network models. Since then, various adversarial attack methods have been proposed in computer vision (Goodfellow, Shlens, and Szegedy 2015; Carlini and Wagner 2017; Madry et al. 2018). Adversarial attacks are not limited to the image domain; they are also possible in domains such as video classification (Wei et al. 2019; Chen et al. 2021b), text classification (Fursov et al. 2021), and speaker recognition (Chen et al. 2021a).

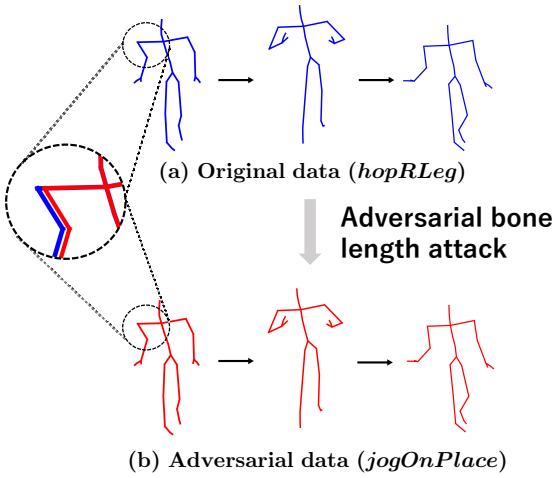


Figure 1: The overview of the adversarial bone length attack. Only the bone length of a skeleton is perturbed (approximately 30 effective dimension). Nevertheless, the classification of the motion changes from (a) *hopRLeg* to (b) *jogOnPlace*.

One of the reasons underlying the current success of adversarial attacks in various domains and tasks is that adversarial attacks manipulate high-dimensional data (Gilmer et al. 2018; Simon-Gabriel et al. 2019). For example, in the image domain, even small datasets (e.g., CIFAR10 Krizhevsky (2009)) have hundreds or thousands of pixels to perturb. In the video domain, more dimensions can be perturbed by exploiting the temporal structure. Recently, it was shown that adversarial attacks can succeed even on data in moderate-dimensional settings (Su, Vargas, and Sakurai 2019; Pony, Naeh, and Mannor 2021). In particular, (Liu, Akhtar, and Mian 2020; Zheng et al. 2020; Wang et al. 2021; Diao et al. 2021) considered adversarial attacks in skeleton-based action recognition. In this case, an attacker is allowed to perturb the shape and motion of the skeleton along frames, and a skeleton at each frame is represented by approximately 100 parameters (two or three-dimensional coordinates of approximately 30 joints). With constraints on

bone connection, the effective number of parameters is even lower. In such a case, it is difficult for attackers to design adversarial perturbations that are sufficiently small to be imperceptible. In (Liu, Akhtar, and Mian 2020; Zheng et al. 2020; Wang et al. 2021; Diao et al. 2021), because the motion of the skeleton along frames also was perturbed, they achieved a high success rate of attack with imperceptible adversarial perturbations.

In this study, we consider an extremely low-dimensional adversarial attack on skeleton-based action recognition, where only the lengths of the skeleton's bones can be perturbed, as shown in Fig. 1. The proposed attack has access to only approximately 30 dimensions, which is significantly lower than existing attacks. Despite the restrictiveness of our new attack setting, we experimentally observed that such an extremely low-dimensional attack can succeed. Our bone length attack only requires people to change the apparent bone lengths, e.g., by covering some parts of the body with clothes or attaching a fake extension to arms. In contrast, existing skeleton-based methods require people to move all the joint positions adversarially, which is almost infeasible because perfect body coordination is required all along the motion. Our experiments were conducted on two datasets, the NTU RGB+D (Shahroudy et al. 2016) and HDM05 (Müller et al. 2007) datasets, and two models, the spatial-temporal graph convolutional network (ST-GCN; Yan, Xiong, and Lin (2018)) and semantics-guided neural network (SGN; Zhang et al. (2020)). In such settings, we saw that with a small perturbation, our attack successfully fooled the models with a 20% success rate, reaching 90% in some cases. We also investigated which parts of a skeleton are more susceptible to our bone length adversarial attack and discovered that perturbing the bones that are long and close to the root joint (base of spine) is effective. Another interesting observation is that adversarial training improves not only the robustness against our attack but also the accuracy of the original data. In the literature (Zhang et al. 2019), it has been established that adversarial training improves the robustness against adversarial examples at the cost of accuracy in original data. Our observations provide a counterexample of this. Based on our experimental results, we consider this to be because in our low-dimensional setting, data augmentation and adversarial examples have similar properties.

Due to the broad and important downstream applications of skeleton-based action recognition, it is crucial to consider the potential vulnerability of skeleton-based action recognition against adversarial attacks. In particular, our adversarial bone length attack seems more realistic than others because it only requires deception of a skeleton extractor to wrongly measure the length of bones, whereas other attacks require perturbation of all positions of joints as well as their time evolution along frames.

Our contributions can be summarized as follows:

- We propose the first adversarial attack that only perturbs the lengths of bones to fool skeleton-based action recognition models. Unlike existing adversarial attacks, our attack works in an extremely low-dimensional setting (approximately 30 dimensions and no temporal perturbation).
- Through extensive experiments, we demonstrate the effectiveness of our bone length attack. We also discovered that bones that are long and close to the base of the spine are significantly more susceptible to the attack than other bones.
- We discovered that adversarial training using the proposed attack improves not only the adversarial robustness against this attack but also the accuracy against the original data. We also observed that data augmentation improves both the clean accuracy and adversarial robustness. This implies that a low-dimensional adversarial attack may have distinct characteristics from other adversarial attacks in high-dimensional settings.

Related Work

Skeleton-based action recognition. Skeleton-based action recognition has attracted significant attention in recent years. There are many advantages to using skeleton data for action recognition. Skeleton data are considered to be robust to lighting, subject clothing, and background (Sun et al. 2020; Wang et al. 2021). They are also superior to RGB data in terms of computational cost. Due to these advantages and the progress of sensors (Zhang 2012) and pose estimation models (Wandt et al. 2021; Xu and Takano 2021), various models have been proposed (Yan, Xiong, and Lin 2018; Zhang et al. 2020; Cheng et al. 2021; Kong, Deng, and Jiang 2021).

Adversarial attacks on skeletons. Adversarial attacks in skeleton-based action recognition have been proposed very recently (Liu, Akhtar, and Mian 2020; Zheng et al. 2020; Wang et al. 2021; Diao et al. 2021). Liu, Akhtar, and Mian (2020) was the first to propose an attack on this task. Their proposed attack generates natural motions satisfying multiple physical constraints by fixing the bone length, limiting the perturbation magnitude and joint acceleration, and using a generative adversarial network (Goodfellow et al. 2014). They also showed that adversarial examples remain adversarial even after being converted from skeleton to RGB data and then from RGB to skeleton data. Zheng et al. (2020) proposed an attack by restricting the change in angle between bones. They also proposed a defense method. Wang et al. (2021) conducted user studies to show that their attack is highly imperceptible to humans. They also argued that joints with high velocity and acceleration are vulnerable features. These attacks assume that the attacker has complete knowledge of the model being attacked, such as its parameters and structure, and are called white box attacks. In contrast, Diao et al. (2021) proposed a black box attack, which assumes complete ignorance of the information about an attacked model. Most of the aforementioned adversarial attacks increase the effective dimensions of the attack to a few thousand by using degrees of freedom in the temporal direction. In contrast, we propose an adversarial attack that has no freedom in the temporal direction. This restriction makes this study largely different from previous studies.

Low-dimensional adversarial attacks. Our proposed attack perturbs only approximately 30 dimensions. There exist

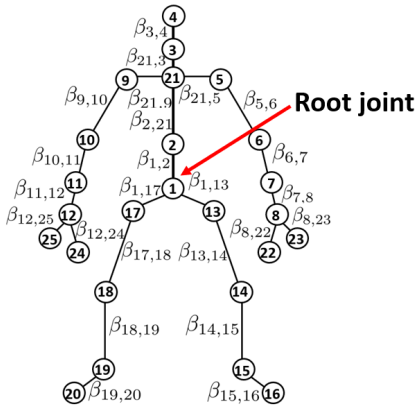


Figure 2: The skeleton of the NTU RGB+D dataset. This skeleton consists of 24 bones, and each joint is indexed. We define the base of the spine (joint 1) as the root joint. Our adversarial bone length attack perturbs the scale parameter $\{\beta_{i,j}\}$.

several related attacks. In the image classification domain, Su, Vargas, and Sakurai (2019) proposed a one-pixel attack, which perturbs only a single pixel. However, attackers can select the target pixel (three channels) from a large number of pixels; thus, the effective dimension is still higher than ours. In addition, a one-pixel attack is relatively easily noticeable. In the video classification domain, the attack proposed by (Pony, Naeh, and Mannor 2021) does not use any spatial information, and thus the available dimensions are low. However, the available dimensions in our attack are less than those in their attack. Moreover, neither Su, Vargas, and Sakurai (2019) nor Pony, Naeh, and Mannor (2021) explored the effect of adversarial training, while we observed an interesting result through adversarial training.

Method

In this section, we explain how to attack skeleton-based action recognition models by perturbing the bone length only.

Notations

We consider L-class skeleton-based action recognition. An action (sequence of skeletons) is represented by $X = \{\boldsymbol{q}_i(t) \mid i = 1, \ldots, M, t = 0, \ldots, T-1\}$, where T and M denote the numbers of frames and joints, respectively, and $\boldsymbol{q}_i(t) \in \mathbb{R}^3$ denotes the 3D coordinates of the i-th joint of the skeleton at frame t. A classifier f receives an action X and outputs an L-dimensional confidence vector, whose k-th entry, denoted by $f(X)_k$, represents the confidence for the k-th class. A class label is encoded in a one-hot representation as $\boldsymbol{y} = (y_1, \ldots, y_L)^\top$.

Bone Length Parameters

Our attack perturbs the lengths of bones. The length of the (i,j)-th bone, which connects the i-th and j-joints, is associated with a scale parameter $\beta_{i,j}$, as shown in Fig. 2. Note

---

Algorithm 1: Pseudocode of adversarial bone length attack

Input: Original sequence of skeleton X, ground truth label y, trained classifier f
Parameter: Step size α, maximum iteration number N, bound ϵ
Output: Adversarial example $\widetilde{X}$

$\boldsymbol{\beta}^{(0)} \leftarrow \mathbf{1}$
for $n \leftarrow 0$ to $N-1$ **do**
  $\widetilde{X} \leftarrow \mathrm{SetParam}\big(\boldsymbol{\beta}^{(n)}, X\big)$
  ▷ $\mathrm{SetParam}(\cdot,\cdot)$ outputs an adversarial example $\widetilde{X}$ consisting of $\widetilde{\boldsymbol{q}}_j(t)$ defined using $\boldsymbol{\beta}^{(n)}$ in Eq. (1).
  $\boldsymbol{z} \leftarrow f(\widetilde{X})$
  ▷ $\boldsymbol{z}$ is the confidence vector, which is output by f that received $\widetilde{X}$.
  $y_{\mathrm{pred}} \leftarrow \mathrm{Predict}(\boldsymbol{z})$
  ▷ $\mathrm{Predict}(\cdot)$ outputs the class label predicted by the classifier f.
  if $y_{\mathrm{pred}} \neq y$ **then**
    break
  end if
  $\boldsymbol{\beta}^{(n+1)} \leftarrow \mathrm{UpdateParam}\big(\boldsymbol{\beta}^{(n)}\big)$
  ▷ $\mathrm{UpdateParam}(\cdot)$ outputs updated bone length parameters via Eq. (3).
  if $n = N-1$ **then**
    $\widetilde{X} \leftarrow \mathrm{SetParam}\big(\boldsymbol{\beta}^{(N)}, X\big)$
  end if
end for
return $\widetilde{X}$

---

that simply perturbing the length scale of each bone does not directly yield a valid skeleton. Therefore, the length scales $\{\beta_{i,j}\}$ of bones are perturbed from the root joint sequentially as follows. First, we define the base of the spine as the root joint $\boldsymbol{q}_{\mathrm{root}}(t)$. Then, starting from the root joint, the position of each joint is set sequentially. Specifically, when a joint $\boldsymbol{q}_i(t)$ is perturbed to $\widetilde{\boldsymbol{q}}_i(t)$, each of its child joints (say, $\boldsymbol{q}_j(t)$) is perturbed as follows:

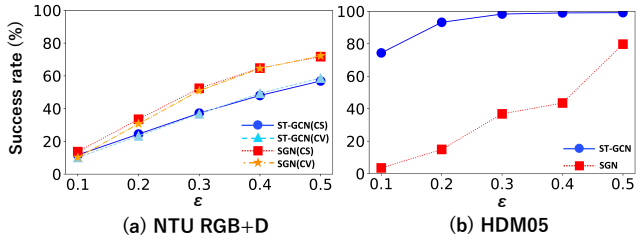
$$\widetilde{\boldsymbol{q}}_j(t) = \beta_{i,j}(\boldsymbol{q}_j(t) - \boldsymbol{q}_i(t)) + \widetilde{\boldsymbol{q}}_i(t). \tag{1}$$

Note that the root is fixed, i.e., $\widetilde{\boldsymbol{q}}_{\mathrm{root}}(t) = \boldsymbol{q}_{\mathrm{root}}(t)$. As can be seen in Eq. (1), the value of $\beta_{i,j}$ is the ratio between the length of the (i,j)-th bone before and after the adversarial perturbation. After the positional rearrangement by Eq. (1), the final adversarial example of X is given by $\widetilde{X} = \{\widetilde{\boldsymbol{q}}_i(t) \mid i = 1, \ldots, M, t = 0, \ldots, T-1\}$.

Adversarial Bone Length Attack

Our attack is based on the following optimization with respect to the bone length parameters $\{\beta_{i,j}\}$. Let $\boldsymbol{\beta}$ be a vector that lists the elements of $\{\beta_{i,j}\}$. Then, the adversarial scales of bones are obtained through the following cross-entropy loss maximization:

$$\max_{\boldsymbol{\beta}} \ \mathcal{L}(\boldsymbol{\beta}) = -\sum_{k=1}^{L} y_k \log f(\widetilde{X})_k, \ \text{s.t.} \ \|\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}\|_\infty \leq \epsilon, \tag{2}$$



(a) NTU RGB+D (b) HDM05

Figure 3: Success rates of the adversarial bone length attack on the ST-GCN and SGN models according to attack strength ϵ: (a) NTU RGB+D; two configurations of the training sets, cross-subject (CS) and cross-view (CV), were considered. Similar results were observed across models and datasets. The SGN model was slightly more vulnerable than the ST-GCN model. (b) HDM05; The ST-GCN model was extremely vulnerable, even for small perturbations.

	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.4$	$\epsilon = 0.5$
ES	51.3%	50.6%	50.4%	50.4%	50.4%
FR	74.6%	81.4%	83.3%	83.1%	82.9%

Table 1: Average of the confidence scores given by the ST-GCN model for misclassified adversarial examples on the HDM05 dataset. The average confidence in early-stopping case (ES) was low, while it was high in full-run case (FR).

where f is the target classifier that we attempt to attack. The initial value of $\boldsymbol{\beta}$ is set to an all-one vector, i.e., $\boldsymbol{\beta}^{(0)} = \mathbf{1}$. The magnitude of the perturbation is bounded by ϵ. We optimize $\boldsymbol{\beta}$ using projected gradient descent (PGD) (Eq. (3)) (Madry et al. 2018). The parameter $\boldsymbol{\beta}$ is iteratively updated by

$$\boldsymbol{\beta}^{(n+1)} = \mathrm{Clip}_{(\boldsymbol{\beta}^{(0)},\epsilon)}\left[\boldsymbol{\beta}^{(n)} + \alpha \cdot \mathrm{sign}(\nabla_{\boldsymbol{\beta}^{(n)}} \mathcal{L}\left(\boldsymbol{\beta}^{(n)}\right))\right], \tag{3}$$

where $\alpha > 0$ is the step size, and $\boldsymbol{\beta}^{(n)}$ is the $\boldsymbol{\beta}$ obtained after n iterations. $\mathrm{Clip}_{(\boldsymbol{\beta}^{(0)},\epsilon)}[\cdot]$ is an operator that clips each entry of a given vector to $[1-\epsilon, 1+\epsilon]$, and $\mathrm{sign}(\cdot)$ denotes the sign function, which maps each element of the argument vector to ± 1 according to its sign.

The pseudo code of our attack is given in Algorithm 1. First, the bone scale parameter $\boldsymbol{\beta}$ and adversarial example $\widetilde{X}$ are initialized as the all-one vector and X, respectively. Then, the following procedure is repeated.

1. $\widetilde{X}$ is input to the target classifier. If the target classifier misclassifies, the attack terminates; otherwise, $\boldsymbol{\beta}$ is updated according to Eq. (3).
2. The action $\widetilde{X}$ is updated according to Eq. (1).

This termination condition of the attack is according to the official code provided by (Wang et al. 2021). In the experiments, we also restricted our attack to a subset of bones to examine which part of skeleton is the most sensitive to the perturbation. In this case, only the subset of $\{\beta_{i,j}\}$ was considered in Eq. (3). We also attacked the classifier using Adam (Kingma and Ba 2015) optimizer instead of PGD, as was used in other studies for attacking skeleton-based action recognition models (Liu, Akhtar, and Mian 2020; Wang et al. 2021). In this case, the update rule in Eq. (3) was replaced with that of Adam.

Experiments and Results

We first introduce our experimental settings. Then, we present the results of our proposed attack. Finally, we attempt to defend against our attack.

Experimental Settings

Datasets We used the NTU RGB+D (Shahroudy et al. 2016) and HDM05 (Müller et al. 2007) datasets, which are 3D skeleton action datasets. The NTU RGB+D dataset consists of 56,880 motion data of a skeleton with 24 bones. This dataset has 60 classes and was created from 40 subjects. There are two ways to split the training and test data: cross-subject (CS) and cross-view (CV). In CS, the dataset is divided so that the numbers of subjects included in the training and test data are both 20. The numbers of samples are 40,320 and 16,560, respectively. CV divides the training and test data according to the camera's viewpoint. The numbers of samples in these training and test datasets are 37,920 and 18,960, respectively. The HDM05 dataset consists of 2,337 data of a skeleton with 30 bones. The number of classes is 130. At the prepossessing stage, we adopted a smoothing filter, translation, and so on. This reduced the number of classes to 65. We randomly divided samples of each class into a training set (80%), validation set (10%), and testing set (10%). Please refer to the supplementary material for this preprocess.

Target models We used two skeleton-based action recognition models as our target models to attack: the ST-GCN (Yan, Xiong, and Lin 2018) and SGN (Zhang et al. 2020) models. When the NTU RGB+D dataset was used, we used pretrained weights provided by the authors of each model[1][2]. When the HDM05 dataset was used, we trained the two models with the official code provided by the authors. To guarantee the convergence, we ran at least 300 epochs for the ST-GCN model and 200 epochs for the SGN model and adopted early stopping for both models. The data augmentation for each model was the same as when the model was trained on the NTU RGB+D dataset. After training, the test accuracies on the ST-GCN and SGN models were 86.1% and 96.1%, respectively.

Evaluation metrics and others Following (Wang et al. 2021), we attacked the test data that were correctly classified by the target model. Then, the attacks were evaluated based on their success rate. In this paper, we used PGD (Eq. 3). We leave the results using Adam in the supplementary material. The maximum number of iterations of the PGD was set to 50. The step size was set to $\alpha = 0.01$, as in (Liu, Akhtar, and Mian 2020). All experiments were conducted using an Intel Core i7-6850K CPU and TITAN RTX GPU.

[1]https://github.com/yysijie/st-gcn
[2]https://github.com/microsoft/SGN

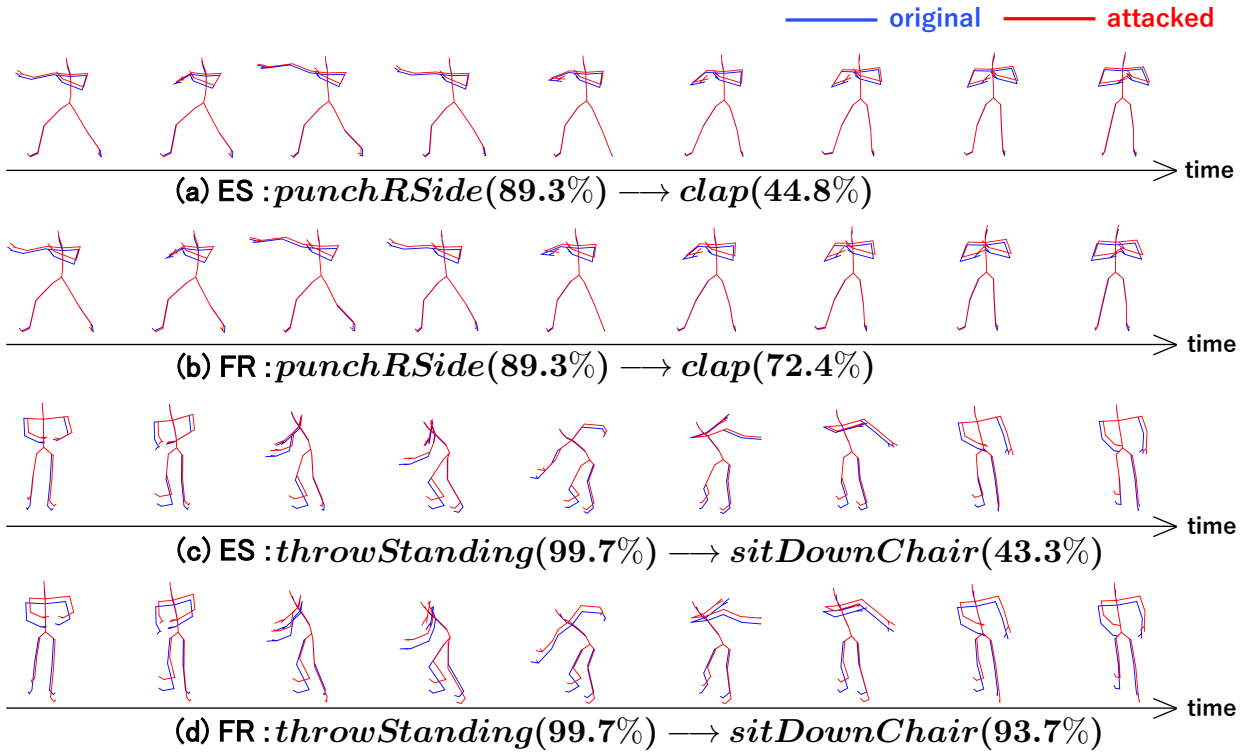


Figure 4: Motion of skeletons before and after adversarial perturbations (blue and red, respectively) when we attacked the ST-GCN model using the HDM05 dataset. In the early-stopping (ES) case, we terminated the iteration in attack once the adversarial example succeeded in fooling the model. In the full-run (FR) case, the adversarial example was updated for a fixed number of 50 iterations. (a, b) the original motion *punchRSide* turned into adversarial motion *clap*. With ES, there is almost no visible change in the skeleton but the prediction confidence remains moderate, whereas with FR, the change becomes slightly more visible and the prediction confidence is high. (c, d) the original motion *throwStanding* turned into adversarial motion *sitDownChair*. The results are similar to those observed in (a,b).

Attack Results

To investigate the effectiveness and imperceptibility of our attack, we attacked two models. In Fig. 3 (a), we show the results of the attack using the NTU RGB+D dataset. As a result of the attack on the ST-GCN model, the success rates for the PDG attack with $\epsilon = 0.1$ and $\epsilon = 0.2$ exceeded 10% and 20%, respectively. Note that this success rate is not as high as those achieved in other studies; this is because the dimensionality of the input data used for the attack in these studies was much higher (e.g., thousands of dimensions), whereas in our setting, the dimensionality is only approximately 30. Nevertheless, our attack achieved moderate success rates that were nonnegligible for practical use. When we set $\epsilon = 0.5$, we can see that the attack became more successful (greater than 50% success). For the HDM05 dataset, the success rate was even higher (Fig. 3 (b)). We observed that the success rate of the attack on the ST-GCN model exceeded 70% when $\epsilon = 0.1$. With $\epsilon = 0.2$, the success rate was over 90%.

The results indicate that the proposed attack is very effective for some models and datasets. In Fig. 4 (a), we provide an adversarial example generated with $\epsilon = 0.1$ when we attacked the ST-GCN model using the HDM05 dataset. One can see that the predicted classes are different although the skeletons appear very similar before and after the attack (blue and red, respectively). The same result was observed for $\epsilon = 0.2$ (Fig. 4 (c)). To summarize, the ST-GCN model trained by the HDM05 dataset was very vulnerable to our bone length attack on the skeleton. However, the confidences on the adversarial motions in Figs. 4 (a) and (c) are relatively low, because we employ early-stopping (ES); the proposed attack terminates once the adversarial example fools the target model (Wang et al. 2021). To make the confidences higher, we can increase the number of iterations in the PGD attack. Hence we attacked the model with 50 iterations, which is the maximum number of iterations in the PGD attack. We call the termination condition the *full-run* (FR) case. Table 1 shows that the average confidence scores in the ES and FR cases for misclassified adversarial examples and that the confidence scores in the FR case are higher than those in the ES case. In Figs. 4 (b) and (d), we demonstrate the FR attacks using the same original data in Figs. 4 (a) and (c), respectively. From Figs. 4, one can see the FR attacks cause bigger changes than the ES attacks. These results

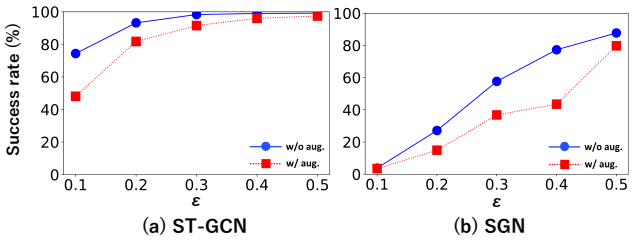


Figure 5: Success rates of the adversarial bone length attack for the (a) ST-GCN and (b) SGN models on the HDM05 dataset with and without data augmentation.

Model	w/o aug.	w/ aug.
ST-GCN	86.1%	89.8%
SGN	95.3%	96.1%

Table 2: Classification accuracy of models trained with the HDM05 dataset with and without data augmentation (*w/ aug* and *w/o aug*, respectively). One can see that data augmentation improved the accuracy for both models.

come from that there is a trade-off between high confidence and imperceptibility.

As we can see the results for the HDM05 dataset shown in Fig. 3(b), attacks on the two models resulted in significantly different success rates; the ST-GCN model was highly vulnerable to the attack, while the SGN model was more adversary-robust. We believe the reason underlying the adversarial robustness of the ST-GCN model lies in the data augmentation process. When we trained the ST-GCN and SGN models with the HDM05 dataset, we followed the training protocol with the NTU RGB+D dataset from the corresponding papers of the two models. Therefore, we did not use data augmentation when the ST-GCN model was trained with the NTU RGB+D dataset, but we did for the SGN model. To observe the effect of data augmentation on adversarial robustness, we trained the models on the HDM05 dataset with and without data augmentation and attacked them. The data augmentation during training of the SGN model was performed by randomly rotating the skeleton. We adopted this augmentation for the ST-GCN model as well because we did not want the results to depend on the quality of the data augmentation. As shown in Table 2, the classification accuracy of both models for original data was improved by data augmentation. Interestingly, as shown in Fig. 5, data augmentation also improved the adversarial robustness (lower success rates). In particular, the improvement in adversarial robustness for the ST-GCN model was significant (approximately 20% improvement at $\epsilon = 0.1$). Therefore, the adversarial robustness exhibited by the ST-GCN model in Fig. 3(b) can be attributed to the presence of data augmentation.

In image classification, some data augmentation methods were proposed to improve adversarial robustness (Zhang et al. 2018; Yun et al. 2019). However, in white box setting, these methods are effective only for non-iterative attack

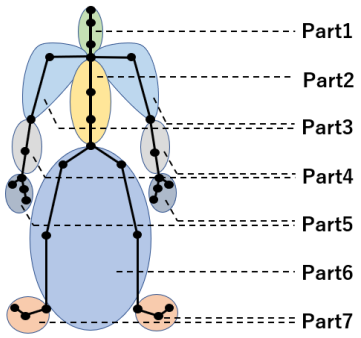


Figure 6: Grouping of bones. The skeleton bones were divided into seven symmetrical parts.

Part	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$
Part 1	0.51%	1.70%	2.20%
Part 2	12.6%	28.9%	40.7%
Part 3	9.34%	25.1%	41.4%
Part 4	2.38%	5.94%	11.9%
Part 5	0.00%	0.51%	0.85%
Part 6	34.1%	67.2%	75.8%
Part 7	0.00%	0.68%	1.35%

Table 3: Success rates of adversarial bone length attack on the ST-GCN model with the HDM05 dataset restricted to different parts. The sensitivity to the attack was strongly dependent on the parts. In particular, Part 6 was the most sensitive to perturbations.

Part	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$
Part3+Part4+Part5	14.4%	39.0%	59.6%

Table 4: Success rates of adversarial bone length attack on the ST-GCN model with the HDM05 dataset by restricted to the union of Parts 3, 4, and 5, which contains more total number of bones as Part 6. The sensitivity to the attack in this region was lower than that in Part 6 (Table 3).

e.g., FGSM (Goodfellow, Shlens, and Szegedy 2015). To the best of our knowledge, the literature has not previously reported data augmentation significantly improving adversarial robustness against iterative attack. In (Rice, Wong, and Kolter 2020) and (Gowal et al. 2020), adversarial training was combined with various data augmentations, but no particular improvement in adversarial robustness was observed. We hypothesize that in the low-dimensional setting of this study, adversarial examples and augmented data are much more similar than in a high-dimensional setting, where adversarial examples are usually considered. We show promising empirical results regarding this hypothesis in the next subsection.

Next, we investigated which subsets of bones were more vulnerable to perturbations. For this, we used the ST-GCN model and the HDM05 dataset. The skeleton was divided into seven parts (Fig. 6), and adversarial perturbation was

only applied to one of them. The results are summarized in Table 3. One can see that attacks on Parts 1, 4, 5, and 7 (head, wrists, hands, and feet, respectively) almost completely failed, whereas attacks on Parts 2, 3, and 6 (body, shoulders, and legs, respectively) were successful. In particular, the success rate for Part 6 was significant compared to those for the other parts. Thus, the legs are more susceptible to attacks. Part 6 has more bones than the other parts. To determine if the number of bones causes the vulnerability, we considered a joint attack on Parts 3, 4, and 5, which have more total bones than Part 6 alone. The results are presented in Table 4. It can be seen that the attack adding perturbations to Parts 3, 4, and 5 resulted in a much lower success rate. As such, the vulnerability of Part 6 is not due to its number of bones.

Based on the above results, we found two tendencies. First, the closer the perturbed bones are to the root joint, the more successful the attack is likely to be. We believe that because a change in bone length can move the positions of the bones of their descendants, a change in bones closer to the root joint can have a greater impact on the entire skeleton. Second, the longer the perturbed bones are, the more likely the attack will be successful. The ϵ in this attack is proportional to the original bone. Therefore, for the same ϵ, the longer the perturbed bone is, the greater the amount of bone change is allowed. These results differ from those of (Wang et al. 2021). They stated that joints with greater velocity and acceleration are more useful for attacks. In contrast, our results show that attacks that perturb the torso are more successful than those that perturb the hands and arms. Therefore, we can see that our attack has different characteristics to theirs.

Adversarial Training

Next, we attempt to defend against our attack by performing adversarial training (Madry et al. 2018) on the HDM05 dataset. We used our attack with $\epsilon = 0.1$ in adversarial training. The results are shown in Fig. 7. One can see that the ST-GCN and SGN models with adversarial training were more robust than that trained using the original data. By adversarial training, the ST-GCN model acquired the same level of robustness against our attack as the SGN model, and the SGN model provided only a low success rate even for large perturbation ϵ. These results suggest that the proposed attack can be prevented to some extent by adversarial training. Notably, we can also see that adversarial training increased the accuracy on the original data, as shown in Table 5. The clean accuracy of the ST-GCN model was 86.1% with standard training and 89.8% with adversarial training. This phenomenon is a counterexample of the following widely seen observation in the literature: adversarial training gains adversarial robustness at the cost of reduced clean accuracy. We speculate that there may be a relationship between our attack and data augmentation as a factor in this phenomenon. As shown in Fig. 5, we achieved some adversarial robustness against our attack by data augmentation. Data augmentation was used to increase the accuracy of the model, similar to our results as shown in Table. 2, and adversarial training with our attack also increased the accuracy. In other words,

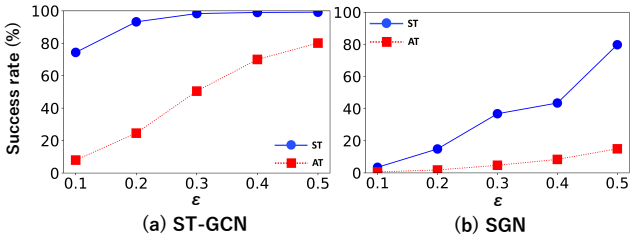


Figure 7: Success rates of the adversarial bone length attack on models with standard training (ST) and adversarial training (AT).

Model	ST	AT
ST-GCN	86.1%	89.8%
SGN	96.1%	95.5%

Table 5: Clean accuracy of models trained with standard training (ST) and adversarial training (AT) on the HDM05 dataset.

adversarial training with our attack has similar properties to data augmentation. The details of this will be the subject of future work.

Conclusion

In this paper, we proposed the first bone length adversarial attack on a skeleton-based action recognition models in an extremely low-dimensional setting. Unlike existing attacks, the proposed attack does not manipulate the motion of skeletons and only perturbs the lengths of the skeleton's bones, the number of which is approximately 30. Nevertheless, for some datasets and settings, it was possible to fool models with small perturbations at a success rate of over 90%. The skeletons before and after the attack appeared very similar, which makes our attack difficult to notice. We also found that perturbing the bones that were longer and closer to the root joint was more effective. Furthermore, we observed some interesting properties, which are considered to be a characteristic of our low-dimensional setting: (i) data augmentation improved both clean accuracy and adversarial robustness, and (ii) adversarial training using our attack also improved both of them. To the best of our knowledge, neither of these results have been reported in the standard high-dimensional setting; thus, we consider that our study opens a new direction for adversarial attacks.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP19K12039.

References

Carlini, N.; and Wagner, D. 2017. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy (SP)*, 39–57.

Chen, G.; Chenb, S.; Fan, L.; Du, X.; Zhao, Z.; Song, F.; and Liu, Y. 2021a. Who is Real Bob? Adversarial Attacks on Speaker Recognition Systems. In *IEEE Symposium on Security and Privacy (SP)*, 694–711.

Chen, Z.; Xie, L.; Pang, S.; He, Y.; and Tian, Q. 2021b. Appending Adversarial Frames for Universal Video Attack. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 3199–3208.

Cheng, K.; Zhang, Y.; He, X.; Cheng, J.; and Lu, H. 2021. Extremely Lightweight Skeleton-Based Action Recognition With ShiftGCN++. *IEEE Transactions on Image Processing*, 30: 7333–7348.

Diao, Y.; Shao, T.; Yang, Y.-L.; Zhou, K.; and Wang, H. 2021. BASAR:Black-Box Attack on Skeletal Action Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7597–7607.

Fursov, I.; Zaytsev, A.; Burnyshev, P.; Dmitrieva, E.; Klyuchnikov, N.; Kravchenko, A.; Artemova, E.; and Burnaev, E. 2021. A Differentiable Language Model Adversarial Attack on Text Classifiers. arXiv:2107.11275.

Gilmer, J.; Metz, L.; Faghri, F.; Schoenholz, S. S.; Raghu, M.; Wattenberg, M.; and Goodfellow, I. 2018. Adversarial Spheres. arXiv:1801.02774.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations (ICLR)*.

Gowal, S.; Qin, C.; Uesato, J.; Mann, T.; and Kohli, P. 2020. Uncovering the Limits of Adversarial Training against Norm-Bounded Adversarial Examples. arXiv:2010.03593.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations(ICLR)*.

Kong, J.; Deng, H.; and Jiang, M. 2021. Symmetrical Enhanced Fusion Network for Skeleton-Based Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(11): 4394–4408.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Liu, J.; Akhtar, N.; and Mian, A. 2020. Adversarial Attack on Skeleton-Based Human Action Recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations (ICLR)*.

Müller, M.; Röder, T.; Clausen, M.; Eberhardt, B.; Krüger, B.; and Weber, A. 2007. Documentation Mocap Database HDM05. Technical Report CG-2007-2, Universität Bonn.

Pony, R.; Naeh, I.; and Mannor, S. 2021. Over-the-Air Adversarial Flickering Attacks Against Video Recognition Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 515–524.

Rice, L.; Wong, E.; and Kolter, J. Z. 2020. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning (ICML)*.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1010–1019.

Simon-Gabriel, C.-J.; Ollivier, Y.; Bottou, L.; Schölkopf, B.; and Lopez-Paz, D. 2019. First-Order Adversarial Vulnerability of Neural Networks and Input Dimension. In *International Conference on Machine Learning (ICML)*, volume 97, 5809–5817.

Su, J.; Vargas, D. V.; and Sakurai, K. 2019. One Pixel Attack for Fooling Deep Neural Networks. *IEEE Transactions on Evolutionary Computation*, 23(5): 828–841.

Sun, Z.; Ke, Q.; Rahmani, H.; Bennamoun, M.; Wang, G.; and Liu, J. 2020. Human Action Recognition from Various Data Modalities: A Review. arXiv:2012.11866.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I. J.; and Fergus, R. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*.

Wandt, B.; Rudolph, M.; Zell, P.; Rhodin, H.; and Rosenhahn, B. 2021. CanonPose: Self-Supervised Monocular 3D Human Pose Estimation in the Wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13294–13304.

Wang, H.; He, F.; Peng, Z.; Shao, T.; Yang, Y.-L.; Zhou, K.; and Hogg, D. 2021. Understanding the Robustness of Skeleton-Based Action Recognition Under Adversarial Attack. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14656–14665.

Wei, X.; Zhu, J.; Yuan, S.; and Su, H. 2019. Sparse Adversarial Perturbations for Videos. *the AAAI Conference on Artificial Intelligence*, 33(01): 8973–8980.

Xu, T.; and Takano, W. 2021. Graph Stacked Hourglass Networks for 3D Human Pose Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16105–16114.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. *the AAAI Conference on Artificial Intelligence*, 32(1).

Yun, S.; Han, D.; Chun, S.; Oh, S. J.; Yoo, Y.; and Choe, J. 2019. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 6022–6031.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations (ICLR)*.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; Ghaoui, L. E.; and Jordan, M. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. In Chaudhuri, K.; and Salakhutdinov, R., eds., *International Conference on Machine Learning (ICML)*, volume 97, 7472–7482.

Zhang, P.; Lan, C.; Zeng, W.; Xing, J.; Xue, J.; and Zheng, N. 2020. Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1109–1118.

Zhang, Z. 2012. Microsoft Kinect Sensor and Its Effect. *IEEE MultiMedia*, 19(2): 4–10.

Zheng, T.; Liu, S.; Chen, C.; Yuan, J.; Li, B.; and Ren, K. 2020. Towards Understanding the Adversarial Vulnerability of Skeleton-based Action Recognition. arXiv:2005.07151.