Understanding the Robustness of Skeleton-based Action Recognition under Adversarial Attack

He Wang[1]*, Feixiang He[1], Zhexi Peng[2], Tianjia Shao[2]†, Yong-Liang Yang[3], Kun Zhou[2], David Hogg[1]
[1]University of Leeds, UK [2]State Key Lab of CAD&CG, Zhejiang University, China
[3]University of Bath, UK
{h.e.wang, scfh, D.C.Hogg}@leeds.ac.uk, {zhexipeng, tjshao, kunzhou}@zju.edu.cn, y.yang@cs.bath.ac.uk

Abstract

*Action recognition has been heavily employed in many applications such as autonomous vehicles, surveillance, etc, where its robustness is a primary concern. In this paper, we examine the robustness of state-of-the-art action recognizers against adversarial attack, which has been rarely investigated so far. To this end, we propose a new method to attack action recognizers which rely on the 3D skeletal motion. Our method involves an innovative perceptual loss which ensures the imperceptibility of the attack. Empirical studies demonstrate that our method is effective in both white-box and black-box scenarios. Its **generalizability** is evidenced on a variety of action recognizers and datasets. Its **versatility** is shown in different attacking strategies. Its **deceitfulness** is proven in extensive perceptual studies. Our method shows that adversarial attack on 3D skeletal motions, one type of time-series data, is significantly different from traditional adversarial attack problems. Its success raises serious concern on the robustness of action recognizers and provides insights on potential improvements.*

1. Introduction

The research in adversarial attack has proven that deep learning is vulnerable to certain imperceptible perturbation on data, leading to security and safety concerns [36]; meanwhile, adversarial attack has been useful in improving the robustness of classifiers [20]. Starting from object recognition, the list of target tasks for adversarial attack has been rapidly expanding, now including face recognition [32], point clouds [45], 3D meshes [47], etc. While adversarial attack on static data (images, geometries, etc.) has been well explored, its effectiveness on time-series has only been attempted under a few settings such as videos [14, 43]. In this paper, we look into another type of time-series data: 3D skeletal motion, for action recognition tasks.

Skeletal motion has been widely used in action recognition [7]. It can greatly improve the recognition accuracy by mitigating issues such as lighting, occlusion and posture ambiguity. In this paper, we show that 3D skeletal motions are vulnerable to adversarial attack but their vulnerability is different from other data. The adversarial attack on 3D skeletal motion faces two unique and related challenges: low redundancy and perceptual sensitivity. When attacking images/videos, it is possible to perturb some pixels without causing too much visual distortion. This largely depends on the redundancy in the image space [37]. Unlike images, which have thousands of Degrees of Freedom (DoFs), a skeletal motion is usually parameterized by fewer than 100 DoFs, i.e. the joints of the skeleton. This not only restricts the space of possible attacks [37], but also affects the imperceptibility of the adversarial samples: a small perturbation on a single joint can be easily noticed. Furthermore, coordinated perturbations on multiple joints in only one frame can hardly work either, because in the temporal domain, similar constraints apply. Any sparsity-based perturbation (on single joints or individual frames) will greatly affect the dynamics (causing jittering or bone-length violations) and will be very obvious to an observer. One consequence is that the perturbation magnitude alone is not anymore a reliable metric to judge the imperceptibility of an attack, as an overall small perturbation could still break the dynamics. This is very different from existing attack tasks where the perturbation magnitude can be heavily relied upon.

To systematically investigate the robustness of action recognizers, we propose a straightforward yet very effective method, Skeletal Motion Action Recognition Attack (SMART), based on an optimization framework that explicitly considers motion dynamics and skeletal structures. The optimization finds perturbations by balancing between classification goals and perceptual distortions, formulated as classification loss and perceptual loss. Varying the classification loss leads to different attacking strategies. The new perceptual loss fully utilizes the dynamics of the motions

*https://youtu.be/DeMkN3efp9s
†Corresponding author

and bone structures. SMART is effective in both white-box and black-box settings, on several state-of-the-art models, across a variety of datasets.

Formally, we systematically investigate the vulnerability of a wide range of state-of-the-art methods under adversarial attack and identify their weaknesses for potential improvements. To this end, we propose a new adversarial attack method with a novel perceptual loss function capturing the perceptual realism and fully exploiting the motion dynamics. We also provide insights into the role of dynamics in the imperceptibility of the adversarial attack based on comprehensive perceptual studies, showing that it is not enough to only constrain the perturbation magnitude, which differs significantly from widely accepted approaches.

2. Related Work

2.1. Skeleton-based Action Recognition

Action recognition is crucial in many applications, namely surveillance, human-robot interaction and entertainment. Recent advances in 3D sensing and pose estimation motivate the use of clean skeleton data to robustly classify human actions, overcoming the biases in raw RGB videos due to body occlusion, scattered background, lighting variation, etc. Unlike conventional approaches that are limited to handcrafted skeletal features [38, 9, 6], recent methods taking the advantage of trained features from deep learning have gained state-of-the-art performance. Based on the representation of skeletal data, deep learning based methods can be classified into three categories, including sequence-based, image-based, and graph-based methods.

Sequence-based methods represent a skeletal motion as a chronological sequence of poses, each of which consists of the coordinates of all the joints. Then RNN-based architecture is employed to perform the classification [7, 23, 35, 53]. Image-based methods represent a skeletal motion as a pseudo-image, which is a 2D tensor where one dimension corresponds to time, and the other dimension stacks all the joints of a single skeleton. Such representation enables CNN-based image classification to be applied to action recognition [24, 16]. Different from the previous two categories that mainly rely on skeleton geometry represented by the joint coordinates, graph-based methods utilize graph representations to naturally consider the skeleton topology (i.e. joint connectivity) which is encoded by bones that connect neighboring joints. Graph neural networks (GNN) are then used to recognize the actions [33, 4, 25, 54, 56]. Based on the code released by the authors, we perform adversarial attacks on the two most representative categories (i.e. RNN- and GNN-based), demonstrating the vulnerability of existing methods.

2.2. Adversarial Attacks

Despite their significant successes, deep neural networks are vulnerable to carefully crafted adversarial attacks as firstly identified in [36]. Delicately designed neural networks with high performance can be easily fooled by unnoticeable perturbations on the input data. With the concern raised, researchers have extensively investigated adversarial attacks on different data types, including 2D images [10, 30, 27, 46, 48], videos [44, 42], 3D shapes [21, 52, 47, 45], physical objects [18, 1, 8], graphs [5], while little attention has been paid to 3D skeletal motions.

The adversarial attack in the context of action recognition is much less explored. Inkawhich et al. [12] perform adversarial attacks on optical-flow based action classifiers, which is mainly inspired by image-based attacks and differs from our work in terms of the input data. The adversarial attack on skeletal motions has just been attempted recently [22, 58] (arXiv only). However, they did not investigate the imperceptibility systematically, which is crucial as shown in our perceptual studies because imperceptibility is a strong requirement on adversarial attack. In our work, we demonstrate better results using a perceptual loss that minimizes the motion derivative deviation relative to the original skeletal motion, thereby preserving the motion dynamics which are intrinsic to actions. This is crucial in attacking highly dynamic motions such as running and jumping. We also perform a perceptual study to systematically validate the imperceptibility of the perturbed skeletal motions and the effectiveness of our choice of perceptual loss.

We demonstrate successful attacks on a range of network architectures, including RNN and GNN based methods, on three datasets. Finally, we present results of three different attacking strategies, including the novel objective of placing the correct action beneath the first n actions in a ranked classification, for a given n.

3. Methodology

SMART is formulated as an optimization problem, where the minimizer is an adversarial sample, for a given motion, that minimizes the perceptual distortion while fooling the target classifier. The optimization has variants constructed for three different attacking strategies: *Anything-but Attack*, *Anything-but-N Attack* and *Specified Attack*. They are used in *white-box* and *black-box* scenarios.

3.1. Optimization for Attack

Given a motion $q = \{q_0, q_1, ..., q_t\}$, where q_t is the frame at time t and consists of stacked 3D joint locations, a trained classifier Φ can predict its class label $y_q = C(\Phi(q))$, where Φ is namely a deep neural network and $\Phi(q)$ is the predicted distribution over class labels. C is usually a *softmax* function and y_q is the predicted label. We aim to find a per-

turbed example, $\hat{q}$, for q, such as $y_q \neq y_{\hat{q}}$. A common method is to find *the minimal perturbation* [49] through solving a constrained optimization. We start with the C&W formulation[2]:

$$\min L_p(q, \hat{q}) \text{ sub. to } C(\Phi(\hat{q})) = c \text{ and } \hat{q} \in [0,1]^n \quad (1)$$

where L_p is a distance function and C is a hard constraint dictating that the predicted class of $\hat{q}$ (bounded in $[0,1]^n$) being c. However, directly solving Eq. 1 is difficult due to that C is highly non-linear [2]. So it can be relaxed by moving the hard constraint into the objective:

$$\text{minimize } L = wL_c(y_{\hat{q}}, c) + (1-w)L_p(q, \hat{q}) \quad (2)$$

where L_c is a classification loss and $w = 0.4$. L_p is normally the perturbation magnitude [2]. But we use a new perceptual loss which is explained later. Eq.2 has intuitive interpretation: there are two forces governing $\hat{q}$. L_c is the classification loss (a relaxed C in Eq.1) where we can design different attacking strategies. L_p is the perceptual loss which dictates that $\hat{q}$ should be visually indistinguishable from q. To optimize for $\hat{q}$, we have only one assumption: we can compute the gradient: $\frac{\partial L}{\partial \hat{q}}$. This way, we can compute $\hat{q}$ iteratively by $\hat{q}_{t+1} = \hat{q}_t + \epsilon f(\frac{\partial L}{\partial \hat{q}_t}, \hat{q}_t)$ where $\hat{q}_t$ is $\hat{q}$ at step t, f computes the updates and ϵ is the learning rate. We set $\hat{q}_0 = q$ and use Adam [17] for f.

3.2. Perceptual Loss

Imperceptibility (governed by L_p in Eq.2) is a hard constraint in adversarial attacks. It requires that human cannot distinguish easily between the adversarial samples and real data. Existing approaches on images and videos achieve imperceptibility by constraining the pixel-wise or frame-wise perturbation magnitude measured by l norms. One major difference in our problem is motion dynamics.

To fully represent the dynamics of a motion, we need the derivatives from *zero-order* (joint location), *first-order* (joint velocity) up to *nth-order*. One common approximation is to use first n terms. When it comes to imperceptibility, the perceived motion naturalness is vital and not all derivatives are at the same level of importance [40]. Inspired by the work in character animation [39, 41, 3], we propose a new perceptual loss:

$$L_p(q, \hat{q}) = \alpha l_{dyn} + (1-\alpha) l_{bl} \quad (3)$$

$$l_{bl} = ||Bl(q) - Bl(\hat{q})||_2^2 = \frac{1}{M}\sum_{i=1}^{M} ||Bl(q_i) - Bl(\hat{q}_i)||_2^2 \quad (4)$$

$$l_{dyn} = \sum_{n=0}^{\infty} \beta_n ||(q^n - \hat{q}^n)||_2^2 \text{ where } \sum_{n=0}^{\infty} \beta_n = 1 \quad (5)$$

where $\alpha = 0.3$. l_{bl} penalizes any bone length deviations in every frame where M is the total frame number. $Bl(q_i) \in \mathbb{R}^{24\times 1}$ is the bone length vector of frame q_i. Theoretically, bone lengths do not change over time. However, they do vary in the original data due to tracking errors. This is why l_{bl} is designed to be frame-wise.

l_{dyn} is the dynamics loss. We use a strategy called *derivative matching*. It is a weighted (by β_n) sum of the l_2 distance between q^n and $\hat{q}^n$, where q^n and $\hat{q}^n$ are the nth-order derivatives and can be computed by forward differencing. Although n goes up to infinity, in practice, we explored up to $n = 4$, which includes joint position, velocity, acceleration, jerk and snap. After exhaustive experiments, we find that enforcing the $0th$, $2nd$ and $4th$ order derivatives while discarding other derivatives gives good results, with the $4th$ derivative adding small gains. Including consecutive derivatives (e.g. $0th$, $1st$ and $2nd$) over-constrains the system. Also, the gain of including higher order derivatives diminishes while incurring more computation. A good compromise is to set $\beta_0 = 0.6$ and $\beta_2 = 0.4$. Matching the $2nd$-order profiles of two motions is critical. For skeletal motions, small location deviations can still generate large acceleration differences, resulting in two distinctive motions. More often, it generates severe jittering and thus totally unnatural motions. An alternative way of regulating the dynamics is to purely smooth the motion, by e.g. minimizing the acceleration. But it dampens highly dynamic motions such as jumping [40]. Also, considering more derivatives above $n = 4$ makes the optimization harder to solve and over-weighs their benefits.

3.3. White-box Attack

With the perceptual loss designed, varying the formulation of the classification loss (L_c in Eq.2) allows us to form different attacking strategies. We present three strategies.

Anything-but Attack (AB) aims to fool the classifier so that $y_q \neq y_{\hat{q}}$. This can be achieved by maximizing the *cross entropy* between $\Phi(q)$ and $\Phi(\hat{q})$:

$$L_c(q, \hat{q}) = -cross\_entropy(\Phi(q), \Phi(\hat{q})) \quad (6)$$

Anything-but-N Attack (ABN) is a generalization of AB. It aims to confuse the classifier so that it has similar confidence levels in multiple classes. ABN is more suitable to confuse classifiers which rely on top N accuracy. In addition, we find that it performs better in black-box attacks by transferability, which will be detailed in experiments. One naive solution is to use multiple AB losses for the top n classes, but it will make the optimization difficult and will not scale as the class number increases. Instead, we propose an easier loss function, maximizing the entropy of the predicted distribution of $\hat{q}$:

$$L_c(q, \hat{q}) = -Entropy(\Phi(\hat{q})), \quad y_q \notin TopN(\Phi(\hat{q})) \quad (7)$$

where $TopN$ is the set of the top n class labels in the predictive distribution $\Phi(\hat{q})$. By minimizing L_c, we actually max-

imize the entropy of $\Phi(\hat{q})$, i.e. forcing it to be flat over all the class labels and thus reduce the confidence of the classifier over any particular class. We stop the optimization once the ground-truth label falls beyond the top n classes. ABN is a harder optimization problem than AB because it needs the predictive distribution to be as flat as possible.

3.3.1 Specified Attack (SA)

Different from AB and ABN, sometimes it is useful to fool the classifier with a pre-defined class label. Given a fake label $y_{\hat{q}}$, we can use its class label distribution $\Phi_{\hat{q}}$, a one-hot vector, and minimize the cross entropy:

$$L_c(q, \hat{q}) = cross\_entropy(\Phi(q), \Phi(\hat{q})) \tag{8}$$

This is the most difficult scenario because it highly depends on the similarity between the source and target label. While turning 'clapping over the head' into 'raising two hands' is achievable with minimal visual changes, turning 'running' into 'squat' without being noticed is much harder.

3.4. Black-box Attack

While the white-box attack relies on the ability to estimate $\frac{\partial L}{\partial \hat{q}}$, which requires the access to the target classifier and is not always possible, black-box attack assumes that the full knowledge of the target classifier is inaccessible. We therefore cannot directly compute $\frac{\partial L}{\partial \hat{q}}$. Under such circumstances, we use *attack-via-transferability* [37]. It begins with training a surrogate classifier. Then adversarial samples are computed by white-box attacks on the surrogate classifier. Finally, the adversarial samples are used to attack the target classifier in a black-box setting. In this paper, we do not construct our own surrogate model. Instead, we use an existing classifier as our surrogate classifier to attack others. In experiments, we attack several state-of-the-art models. To test the transferability and generalizability of our method, we use every model in turn as the surrogate model and attack the others.

4. Experimental Results

We first introduce the datasets and models for our experiments, followed by our white-box and black-box results. We then present our perceptual studies on the imperceptibility and compare SMART with other methods. During the attack, we first use the source code shared by the authors if available or implement the methods ourselves. Then we train them strictly following the protocols in their papers. Next, we test the models and collect the data samples that the trained classifiers can successfully recognize, to create our adversarial attack datasets. Finally, we compute the adversarial samples using different attacking strategies.

4.1. Datasets

We choose three widely used datasets. **HDM05** [28] contains 2337 sequences for 130 actions performed by 5 non-professional actors. The 3D joint locations of the subjects are provided in each frame. **MHAD** [29] is captured using a multi-modal acquisition system, consisting of 11 actions performed by 12 subjects, where 5 repetitions are performed for each action, resulting in 659 sequences. In each frame, the 3D joint positions are extracted based on the 3D marker trajectories. **NTU60** [31] is captured by Kinect v2 and is currently one of the largest publicly available datasets for 3D action recognition. It is composed of more than 56,000 action sequences. A total of 60 action classes are performed by 40 subjects. The 3D coordinates of joints are provided by Kinect. Due to the huge number of samples and the large intra-class and viewpoint variations, the NTU60 is very challenging and is highly suitable to validate the effectiveness and generalizability of our approach. Note we exclude Kinectics [15], a dataset that is also used in many papers, for two reasons. First, some older recognizers we investigate cannot achieve reasonable classification accuracy on it. Second, its quality is too low to evaluate the success of the attack, explained in Section 4.5.

4.2. Target Models

Rather than focusing only on the most recent methods, we select a range of methods: HRNN [51], ST-GCN [50], AS-GCN [19], DGNN [33], 2s-AGCN [34], MSG3D [26] and SGN [55], and investigate their vulnerability under different scenarios. They include both RNN- and GNN-based models. We implement HRNN following the paper and use the code shared online for the rest of the methods. We also follow their protocols in data pre-processing. Specifically, we preprocess the HDM05 and MHAD as in [51] (where HDM05 is grouped into 60 classes), and the NTU60 as in [34]. We also map different skeletons to a standard 25-joint skeleton as in [40].

4.3. White-box Attack

In this section, we qualitatively and quantitatively evaluate the performance of SMART. We use a learning rate between 0.005 and 0.0005 and a maximum of 300 iterations. The setting for AB and ABN is straightforward. In SA, the number of experiments needed would be prohibitively large if we were to attack every motion with every other label but the ground-truth. Instead, we randomly select fake labels to attack. Since the number of motions attacked is large, the results are sufficiently representative. Note that this is a very strict test as most of the motions are rather distinctive. For simplicity, we only show representative results in the paper. For more results, please refer to the supplementary materials and video.

4.3.1 Attack Results.

We show the quantitative results of AB in Table 1 Left. High success rates are universally achieved across different datasets and target models, demonstrating the generalizability of SMART. For adversarial attack, it is not surprising if the before-attack and after-attack labels are semantically similar, e.g. from drinking water to eating. In SMART, a variety of examples are found where the after-attack labels are significantly different from the original ones. Due to the space limit, we leave all the details in the supplementary video and materials and only give a couple of examples here. In HDM05, high confusion is found between turn_L (turn left) and walk_rightRC (walk sideways, to the right, feet cross over alternately front/back) in HRNN. Similarly, in NTU, high confusion is found between standing_up (from sitting) and wear_a_shoe in 2SAGCN. These labels have completely different semantics and involve different body parts and motion patterns. Moreover, this kind of confusion is observed across all datasets and models.

We show the ABN results in Table 1 Mid, in two variations: AB3 and AB5, as a generalization of AB. They are good for attacking classifiers based on top N accuracy. ABN is a harder problem than AB, with AB5 being harder than AB3, hence has a lower success rate. In terms of datasets, MHAD is the hardest for ABN because there are only 11 classes as opposed to 65 and 60 in the other two. Excluding the ground-truth label from the top 5 out of 11 classes is much more challenging than that of 65 and 60 classes.

Table 1 Right shows the SA results. SA is the most difficult because randomly selected class labels often come from significantly different action classes. Although it might be easy to confuse the model between 'deposit' and 'grab', it is extremely difficult to do so for 'jumping' and 'wear-a-shoe'. However, even under such circumstances, SMART is still able to succeed in more than 70% cases on average, with multiple tests above 96% and even achieving 100%.

Performance. The major computational cost comes from the gradient estimation which depends on the target model because it requires back-propagation. We run a maximum of 300 iterations. The total amount of time each iteration takes are on average 0.102s, 0.267s, 0.419s, 0.275s and 0.738s on HRNN, ST-GCN, AS-GCN, DGNN and 2S-AGCN respectively, on Nvidia GTX 1080Ti (DGNN and 2S-AGCN) and TitanXp (HRNN, ST-GCN and AS-GCN).

4.4. Black-box Attack

In the black-box setting, we attack the NTU dataset. Since we need a surrogate model to fool the target models, we first use 2s-AGCN as the surrogate model to attack DGNN, AS-GCN, MSG3D and SGN. The results are shown in Table 2. We notice that SMART achieves successes on all target models except MSG3D, which indicates that not all target models are equally easy to fool by the transferred black-box attack. To further investigate it, we use three models: AS-GCN, DGNN and 2s-AGCN, and in turn take every model as the surrogate model and produce adversarial examples using AB and AB5.

Results are shown in Table 3. AB5 results are in general better than AB. We speculate that there are two factors. First, the predictive class distribution of AB5 is likely to be flatter than AB. The flatness improves the transferability because a target model with similar decision boundaries will also produce a similarly flat predictive distribution, and thus is more likely to be fooled. Besides, since the ground-truth label is pushed away from the top 5 classes in the surrogate model, it is also likely to be far away from the top in the target model. We also notice that the transferability is not universally successful. DGNN and AS-GCN cannot easily fool one another. Meanwhile, 2S-AGCN can fool and be fooled by both of them. Since the transferability can be described by distances between decision boundaries [37], our speculation is that 2S-AGCN's boundary structure overlaps with both DGNN and AS-GCN significantly but the other two overlap little. The theoretical reason is hard to identify, as the formal analysis on transferability has just emerged on static data [37, 57]. The theoretical analysis of time-series data is beyond the scope of this paper and is therefore left for future work.

4.5. Perceptual Study

One key difference between SMART and existing work is that we employ both *numerical accuracy* and *rigorous perceptual studies* to evaluate the success of attacks. Imperceptibility is a requirement for any adversarial attack. All the success shown above would have been meaningless if the attack were noticeable to humans. To evaluate imperceptibility, qualitative visual comparisons can be used on the image-based attack, but rigorous perceptual studies are needed for complex data [47], as the numerical success can always be achieved by sacrificing the imperceptibility. This is especially the case for motions. Also, the necessity of perceptual studies restricts us from using noisy datasets (e.g. Kinetics [15]) because the subjects are unable to identify perturbations in side-by-side comparisons due to the excessive jittering and tracking errors in the original data.

We conduct three user studies (Deceitfulness, Naturalness and Indistinguishability). Since our sample space is huge (7 models $\times$ 3 datasets $\times$ 3 attacking strategies), we choose the most representative setting. We use the adversarial samples under AB in HDM05 and MHAD. NTU dataset is only used in visual evaluation, not perceptual study due to motion jittering in the original data (see the video for details). In total, we recruited totally 41 subjects (age between 18 and 37). Details are in the supplementary materials.

Deceitfulness. In each user study, we randomly choose 100 motions with the ground-truth and the after-attack la-

Model/Data	HDM05	MHAD	NTU	HDM05	MHAD	NTU	HDM05	MHAD	NTU
HRNN	100	100	99.56	100/100	100/100	99.84/99.62	67.19	57.41	49.17
ST-GCN	99.57	99.96	100	93.30/90.28	76.86/70.5	95.86/91.32	74.95	66.93	100
AS-GCN	99.36	92.84	97.43	91.46/82.83	42.07/22.34	91.18/82.47	64.62	40.18	99.48
DGNN	96.09	94.46	92.51	93.55/86.32	87.54/74.27	98.73/97.62	97.26	96.13	99.99
2s-AGCN	99.18	95.97	100	83.40/75.2	55.9/32.08	100/100	96.72	97.53	100
mean	98.84	96.65	97.9	92.34/86.93	72.47/59.84	97.12/94.21	80.15	71.64	89.73

Table 1. Success rate. Left: Anything-but (AB) Attack. Mid: Anything-but-N Attack. The results are AB3/AB5 when n = 3 (AB3) and 5 (AB5). Right: Specified Attack (SA).

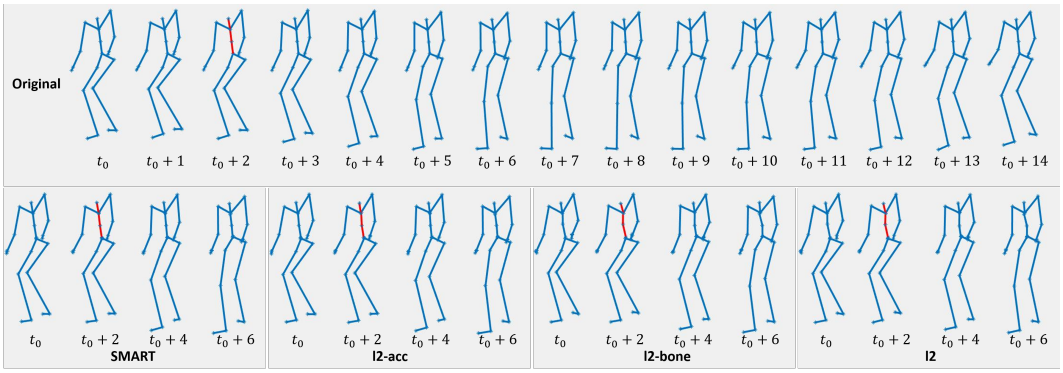

Figure 1. Visual comparison between different losses. Highlighted spine areas in the same frame show key visual differences.

DGNN	AS-GCN	MSG3D	SGN
98.37	98.10	3.08%	97.75%

Table 2. Success rate of AB black-box attack, using 2s-AGCN.

	DGNN	2s-AGCN	AS-GCN
DGNN	n/a	90.6(90.99)	7.24(7.63)
2s-AGCN	98.37(98.46)	n/a	98.10(98.96)
AS-GCN	10.90(12.97)	91.17(91.99)	n/a

Table 3. Success rate (AB/AB5) of black-box attack.

bel for 100 trials. In each trial, the video is played for 6 seconds and then the subject is asked to choose which label best describes the motion with no time limit. This is to test whether SMART visually changes the semantics of the motion. This is also to test whether people can distinguish actions by only observing skeletal motions.

Naturalness. Since unnatural motions can be easily identified as a result of the attack, we perform ablation tests on different loss term combinations. We design four settings: l2, l2-acc, l2-bone, SMART. l2 is where only the l_2 norm of joint perturbation is used, which is also widely used in existing methods such as image/video/mesh attack. l2-acc is l2 plus the acceleration loss, l2-bone is l2 plus the bone-length loss and SMART is the proposed perceptual loss. We first show static poses in Figure 1. Motion comparisons are available in the supplementary video. Visually, SMART is the best. Even from static poses, one can easily see the artifacts caused by joint displacements. The spinal joints are the most obvious. The joint displacements cause unnatural zig-zag bending in l2, l2-acc and l2-bone, which is even more obvious in motions.

Next, we conduct perceptual studies. In each study, we randomly select 50 motions. For each motion, we make two trials. The first includes one attacked motion by SMART and one randomly selected from l2, l2-acc and l2-bone. The second includes two motions randomly drawn from l2, l2-acc and l2-bone. The first trial evaluates our results against other alternatives and the second reveals the impact of different perceptual loss terms. In each of the 100 trials, two motions are played together for 6 seconds twice, and then the subject is asked to choose which motion looks more natural or cannot tell the difference, with no time limit.

Indistinguishability. In this study, we conduct a very strict test to see if the users can tell if a motion is perturbed in any way at all. In each experiment, 100 pairs of motions are randomly selected. In each trial, the left motion is always the original and the user is told so. The right one can be the original (**sensitivity**) or attacked (**perceivability**). We ask if the user can see any visual differences. Each video is played for 6 seconds then the user is asked to choose if the right motion is a changed version of the left, with no time limit. This user study serves two purposes. Perceivability is a direct test on Indistinguishability on the attack while sensitivity is to screen out subjects who tend to give random choices. Most users are able to recognize if two motions are the same (close to 100% accuracy), but there are a few whose choices are more random. We discard any user data which falls below 80% accuracy on the

sensitivity test.

4.5.1 Results.

The success rate of **Deceitfulness** is 93.32% overall, which means that most of the time SMART does not visually change the semantics of the motions. When looking into the success rate on different datasets, SMART achieves 86.77% on HDM05 and 96.38% on MHAD. This also shows that most of the time people can tell different actions by observing skeletal motions, even for similar actions. Next, Figure 2 Left shows the results of **Naturalness**. Users' preferences over different losses are SMART $>$ l2-acc $>$ l2 $>$ l2-bone. SMART leads to the most natural results as expected.

Finally, we conduct the **Indistinguishability** test. The final results are 81.9% on average, 80.83% on HDM05 and 83.97% on MHAD. Note that this is a side-by-side comparison and thus is very harsh. The users are asked to find any visual differences. To avoid situations where motions are too fast to spot any differences (e.g. kicking and jumping motions), we also play the motions three times more slowly than the original. Even under such harsh tests, humans still cannot spot any difference most of the time.

4.6. Classifier Robustness under SMART Attack

After rigorously confirming the effectiveness of SMART across datasets and models, we analyze the results to investigate the vulnerability of the target models. We start by looking at which joint or joint groups are attacked the most. Initially, if some joints tend to be attacked together, the correlations between the joint perturbations should be high. So we compute the Pearson correlations of joint perturbations, shown in Figure 3 Left. Although some local high correlations can be found (e.g. between joint 2 and 3, 6 and 7, 9 and 10, 20 and 21), they are not universal. Please see other results in the supplementary material. Next, we assume that the attack behavior might be class-dependent, i.e. depending on actions. However, after computing the joint perturbation correlations based on actions, no consistent and obvious patterns is found either.

Finally, we find that the displacement-speed and displacement-acceleration correlations reveal a consistent description of the vulnerability, shown in Figure 3 Mid and Right. The correlations are computed between the joint displacements and the original velocities and accelerations, respectively. These two correlations reveal the joint vulnerability: the higher the speed/acceleration is, the more the joint is attacked (shown by the high values along the main diagonal). In addition, they also reveal some consistent across-joint correlations (as shown by red boxes). Note that the joints in a red box belong to one part of the body (four limbs and one trunk). These joints normally have high within-group correlations in motions. Coordinated attacks on them easily fool the action recognizers.

The analysis suggests that joints with high velocity and acceleration are important features in the target models because these joints are attacked the most. This is especially so for joint groups with high within-group correlations. Most of the tested models are very sensitive to perturbations to these features, raising a big concern. Meanwhile, the analysis also suggests that reducing the sensitivity of a classifier over these features will increase its resistance to adversarial attack. To this end, one possible solution is to induce noises around the perturbation gradient during training, instead of purely white noises used by many methods. Another possibility is to introduce semantic descriptors (e.g. featuring a waving motion as one hand moving side-to-side above the head) which are not sensitive to small changes in these raw features.

Dynamics in Attack Imperceptibility. To investigate the role of dynamics compared with joint-only perturbation, we conduct further analysis on SMART-vs-l2 where users prefer SMART to l2. We first compute their respective joint-wise deviations from the original motions, shown in Figure 2 Right. In general, the perturbations of SMART are in general higher than l2 and have larger standard deviations. However, the users still choose SMART over l2. It indicates that with proper exploitation of dynamics, larger perturbations can generate even more desirable results. This is somewhat surprising and significantly different from the static data (e.g. images), where it is believed that the perturbation magnitude is tightly tied to imperceptibility [11]. This also suggests that classifiers could use perturbations on the dynamics to make the training more robust, which is complementary to the afore-mentioned suggestion of inducing noises around the perturbation gradient.

4.7. Comparison

To show that SMART is an effective tool for attack analysis, we compare SMART with IAA [13] and CIASA [22]. As there are two competing factors (attack success vs imperceptibility), we fix one and compare the other. The success rate is largely governed by the clipping threshold of the perturbation magnitude in IAA and CIASA, and is hence easily tunable, while user studies on imperceptibility are expensive. We, therefore, tune IAA & CIASA to achieve similar success rates, then conduct perceptual studies for comparison. Specifically, we conduct AB attack on HDM05 and the Indistinguishability test, as AB is also used in both papers. Each experiment includes 120 pairs of motions including motions evenly sampled from the original motions, SMART, IAA and CIASA results (30 motions each). In each trial, the left motion is the original motion while the right one is either the original motion, a SMART sample, an IAA sample or a CIASA sample. Results are shown in Table 4. While the attack success rates of the three methods

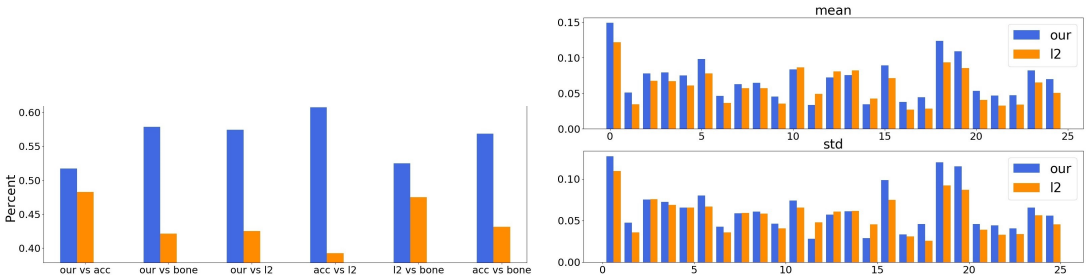

Figure 2. Left: Normalized user preference on Naturalness. our: SMART. bone: l2-bone. acc: l2-acc. The vertical axis is the percentage of user preference. Right: The mean (Top) and standard deviation (Bottom) of the joint-wise deviations of SMART and l2.

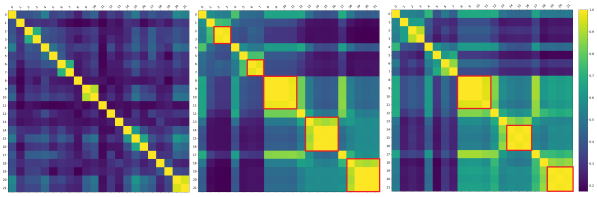

Figure 3. 2S-AGCN on HDM05, displacement-displacement correlations (Left), displacement-speed correlations (Middle) and displacement-acceleration correlations (Right).

Model/Method	SMART	IAA	CIASA
HRNN	100%	98.12%	98.75%
STGCN	99.57%	99.57%	99.56%
2S-AGCN	99.18%	98.77%	98.98%
HRNN	**42.22%**	36.67%	32.22%
STGCN	**90.00%**	87.5%	**90.00%**
2S-AGCN	**80.83%**	35.33%	49.33%

Table 4. Success rate in attack (Upper) and Indistinguishability (Lower). The attack success rate is the best results for SMART, IAA and CIASA.

are similar, SMART, in general, generates more indistinguishable adversarial samples than IAA and CIASA do. We notice that most failures of IAA and CIASA are caused by broken motion dynamics and are therefore easily perceivable. This is understandable because IAA does not consider dynamics and thus generates jittering motions; CIASA uses GANs to govern the motion quality, which can only generate plausible motions, but not imperceptible samples. Details can be found in the supplementary materials.

5. Discussion

Imperceptibility is vital in adversarial attack. When it comes to skeletal motions, perceptual studies are essential because there is no widely accepted metrics that fully reflect perceived realism/naturalness/quality. In addition, it helps us to uncover a unique feature of attacking skeletal motions. Losses solely based on perturbation magnitude are often overly conservative because they are mainly designed for attacking static data and unable to fully utilize the dynamics. Next, forming the joint deviation as a hard constraint [22] via clipping is not the best strategy. The threshold needs to be manually tuned and it varies based on data. Besides, our perceptual study shows that larger perturbations can be used if the dynamics are exploited properly.

SMART is a straightforward but surprisingly effective attack method across datasets, models, attack strategies, and harsh perceptual studies. The simplicity of SMART raises an alarming concern for current action recognition research as it does not require complex computation to attack the state-of-the-art models. Through analysing SMART's behavior, we identified one key cause of their vulnerability: the over-sensitivity to joints with high velocity and acceleration, which we hope will help the future research to improve the recognition robustness.

6. Conclusion and Future Work

We demonstrated the vulnerability of several state-of-the-art action recognizers under adversarial attack. To this end, we proposed a new method, SMART, to attack action recognizers based on 3D skeletal motions. Through comprehensive qualitative and quantitative evaluations, we showed that SMART is *general* across multiple state-of-the-art models on various benchmark datasets. Moreover, SMART is *versatile* since it can deliver both white-box and black-box attacks with multiple attacking strategies. Finally, SMART is *deceitful* as verified in extensive perceptual studies. Based on SMART, we revealed possible causes of the vulnerability of several state-of-the-art models. In the future, we would like to theoretically investigate why the transferability varies between different models under the black-box attack. We will also investigate how to systematically resist adversarial attack.

Acknowledgements: We thank Qun-Ce Xu and Kai-Wen Hsiao for their help on the perceptual study. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 899739 CrowdDNA, EPSRC (EP/R031193/1), NSF China (No. 61772462, No. U1736217), RCUK grant CAMERA (EP/M023281/1, EP/T014865/1) and the 100 Talents Program of Zhejiang University.

References

[1] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. *arXiv*, abs/1707.07397, 2017. 2

[2] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017. 3

[3] Wenheng Chen, He Wang, Yi Yuan, Tianjia Shao, and Kun Zhou. Dynamic future net: Diversified human motion generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 2131–2139, New York, NY, USA, 2020. Association for Computing Machinery. 3

[4] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[5] Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. Adversarial attack on graph structured data. In *Proceedings of the 35th International Conference on Machine Learning*, 2018. 2

[6] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo. 3d human action recognition by shape analysis of motion trajectories on riemannian manifold. *IEEE Transactions on Cybernetics*, 2015. 2

[7] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, 2015. 1, 2

[8] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on machine learning models. *arXiv*, abs/1707.08945, 2017. 2

[9] B. Fernando, E. Gavves, M. José Oramas, A. Ghodrati, and T. Tuytelaars. Modeling video evolution for action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5378–5387, 2015. 2

[10] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2014. 2

[11] Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge J. Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. *CoRR*, abs/1907.10823, 2019. 7

[12] Nathan Inkawhich, Matthew Inkawhich, Yiran Chen, and Hai Li. Adversarial attacks for optical flow-based action recognition classifiers. *arXiv*, abs/1811.11875, 2018. 2

[13] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. Muller. Adversarial attacks on deep neural networks for time series classification. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019. 7

[14] F. Karim, S. Majumdar, and H. Darabi. Adversarial attacks on time series. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 1

[15] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 4, 5

[16] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid. A new representation of skeleton sequences for 3d action recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4570–4579, 2017. 2

[17] Diederik Kingma and Jimmy Ba. Adam : A method for stochastic optimization. *arXiv*, abs/1412.6980v9, 2014. 3

[18] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv*, abs/1607.02533, 2016. 2

[19] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 4

[20] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[21] Hsueh-Ti Derek Liu, Michael Tao, Chun-Liang Li, Derek Nowrouzezahrai, and Alec Jacobson. Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer. In *International Conference on Learning Representations*, 2019. 2

[22] Jian Liu, Naveed Akhtar, and Ajmal Mian. Adversarial attack on skeleton-based human action recognition. *arXiv*, abs/1909.06500, 2019. 2, 7, 8

[23] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 816–833, 2016. 2

[24] Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recogn.*, 68(C):346–362, 2017. 2

[25] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[26] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4

[27] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, 2016. 2

[28] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database hdm05. Technical Report CG-2007-2, Universität Bonn, 2007. 4

[29] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Berkeley mhad: A comprehensive multimodal human action database. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 53–60, Jan 2013. 4

[30] Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. *arXiv*, abs/1511.07528, 2015. 2

[31] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016. 4

[32] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security*, 2016. 1

[33] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 7912–7921, June 2019. 2, 4

[34] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 4

[35] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 2

[36] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Synthesizing robust adversarial examples. *arXiv*, abs/1312.6199, 2014. 1, 2

[37] Florian Tramèr, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. The space of transferable adversarial examples. *arXiv*, abs/1704.03453, 2017. 1, 4, 5

[38] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2014. 2

[39] He Wang, Edmond SL Ho, and Taku Komura. An energy-driven motion planning method for two distant postures. *IEEE transactions on visualization and computer graphics*, 21(1):18–30, 2015. 3

[40] H. Wang, E. S. L. Ho, H. P. H. Shum, and Z. Zhu. Spatio-temporal manifold learning for human motions via long-horizon modeling. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1, 2019. 3, 4

[41] He Wang, Kirill A Sidorov, Peter Sandilands, and Taku Komura. Harmonic parameterization by electrostatics. *ACM Transactions on Graphics (TOG)*, 32(5):155, 2013. 3

[42] Jue Wang and Anoop Cherian. Learning discriminative video representations using adversarial perturbations. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV*, 2018. 2

[43] Xingxing Wei, Jun Zhu, and Hang Su. Sparse adversarial perturbations for videos. *arXiv*, abs/1803.02536, 2018. 1

[44] Xingxing Wei, Jun Zhu, Sha Yuan, and Hang Su. Sparse adversarial perturbations for videos. In *AAAI*, 2018. 2

[45] Chong Xiang, Charles Qi, and Bo Li. Generating 3d adversarial point clouds. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9136–9144, June 2019. 1, 2

[46] Chaowei Xiao, Bo Li, Jun yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. In *IJCAI*, pages 3905–3911, 2018. 2

[47] Chaowei Xiao, Dawei Yang, Bo Li, Jia Deng, and Mingyan Liu. Meshadv: Adversarial meshes for visual recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6898–6907, 2019. 1, 2, 5

[48] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *International Conference on Learning Representations*, 2018. 2

[49] Han Xu, Yao Ma, Haochen Liu, Debayan Deb, H. S. Liu, Jiliang Tang, and Anil Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17:151–178, 2020. 3

[50] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 4

[51] Yong Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, June 2015. 4

[52] Xiaohui Zeng, Chenxi Liu, Yu-Siang Wang, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi-Keung Tang, and Alan L. Yuille. Adversarial attacks beyond the image space. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4302–4311, 2019. 2

[53] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng. View adaptive neural networks for high performance skeleton-based human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1963–1978, 2019. 2

[54] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[55] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4

[56] Xikun Zhang, Chang Xu, and Dacheng Tao. Context aware graph convolution for skeleton-based action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[57] Chenxiao Zhao, P. Fletcher, Mixue Yu, Yaxin Peng, Guixu Zhang, and Chaomin Shen. The adversarial attack and detection under the fisher information metric. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 5

[58] Tianhang Zheng, Sheng Liu, Changyou Chen, Junsong Yuan, Bangmin Li, and Kui Ren. Towards understanding the adversarial vulnerability of skeleton-based action recognition. *ArXiv*, abs/2005.07151, 2020. 2