Understanding the vulnerability of skeleton-based Human Activity Recognition via black-box attack

Yunfeng Diao [a,1], He Wang [b,*,1], Tianjia Shao [c], Yongliang Yang [d], Kun Zhou [c], David Hogg [e], Meng Wang [a]

[a] *School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China*
[b] *Department of Computer Science, University College London, London, UK*
[c] *State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China*
[d] *Department of Computer Science, University of Bath, Bath, UK*
[e] *School of Computing, University of Leeds, Leeds, UK*

ARTICLE INFO

ABSTRACT

Human Activity Recognition (HAR) has been employed in a wide range of applications, e.g. self-driving cars, where safety and lives are at stake. Recently, the robustness of skeleton-based HAR methods have been questioned due to their vulnerability to adversarial attacks. However, the proposed attacks require the full-knowledge of the attacked classifier, which is overly restrictive. In this paper, we show such threats indeed exist, even when the attacker only has access to the input/output of the model. To this end, we propose the very first black-box adversarial attack approach in skeleton-based HAR called BASAR. BASAR explores the interplay between the classification boundary and the natural motion manifold. To our best knowledge, this is the first time data manifold is introduced in adversarial attacks on time series. Via BASAR, we find on-manifold adversarial samples are extremely deceitful and rather common in skeletal motions, in contrast to the common belief that adversarial samples only exist off-manifold. Through exhaustive evaluation, we show that BASAR can deliver successful attacks across classifiers, datasets, and attack modes. By attack, BASAR helps identify the potential causes of the model vulnerability and provides insights on possible improvements. Finally, to mitigate the newly identified threat, we propose a new adversarial training approach by leveraging the sophisticated distributions of on/off-manifold adversarial samples, called mixed manifold-based adversarial training (MMAT). MMAT can successfully help defend against adversarial attacks without compromising classification accuracy.

1. Introduction

Human Activity Recognition (HAR) solutions are mainly based on deep learning, which are vulnerable to adversarial attack [1]. This causes major concerns especially in safety and security [2], as the perturbations are *imperceptible* to humans but *destructive* to machine intelligence. Detecting and defending against attacks have been actively investigated [2]. While the research on static data (e.g. images, texts, graphs) has been widely studied, the attack on time-series data has only been recently explored [3]. We investigate a specific yet important type of time series data, skeletal motions, in HAR.

Skeletal motion is widely employed in HAR to mitigate issues such as lighting, occlusion, view angles, etc. Therefore, the vulnerability of skeleton-based classifiers under adversarial attack has recently drawn attention [4–6]. Albeit identifying a key issue that needs to be addressed, their methods are essentially *white-box*. The only attempt on

black-box attack is via surrogate models, i.e. attack a classifier in a white-box manner then use the results to attack the target classifier. While white-box attack requires the full knowledge of the attacked model which is overly restrictive, black-box attack via surrogate models cannot guarantee success due to its heavy dependence on the choice of the surrogate model [6]. *Is true black-box attack possible in skeleton-based HAR?* To answer the question, we restrict the accessible knowledge to be only the inputs/outputs of the classifiers, and propose BASAR, the *very first* black-box attack method on skeleton-based activity recognition to our best knowledge.

Black-box attack on skeletal motions brings new challenges due to their unique features compared with other data. First, a skeleton usually has less than 100 Degrees of freedom (Dofs), much smaller than previously attacked data such as images/meshes. This low dimensionality leads to low-redundancy, restricting possible attacks within small

---

subspaces. Second, *imperceptibility* is a prerequisite for any successful attack, but its evaluation on skeletal motions is under-explored. Different from the attack where visual imperceptibility has high correlations with the perturbation magnitude (e.g. images), a skeletal motion has dynamics that are well-recognized by human perception. Any sparse attack, e.g. on individual joints or individual poses, albeit small, would break the dynamics and therefore be easily perceptible. In contrast, coordinated attacks on all joints and poses can provide better imperceptibility even when perturbations are relatively large [6]. Consequently, the perturbation magnitude alone (as in most existing methods) is not a reliable metric for skeletal motion. Last but not least, prior methods assume that adversarial samples are off the data manifold [7]. As we will show, skeletal motion is one real-world example where on-manifold adversarial samples not only exist but are rather common, raising serious concerns as these on-manifold adversarial samples are *implementable*.

Given a motion $\mathbf{x}$ with class label $c_{\mathbf{x}}$, BASAR aims to find $\mathbf{x}'$ that is close to $\mathbf{x}$ (measured by some distance function) and can fool a black-box classifier such that $c_{\mathbf{x}'} \neq c_{\mathbf{x}}$. BASAR formulates this process as a constrained optimization problem, aiming to find $\mathbf{x}'$ that is just outside $c_{\mathbf{x}}$ with a new requirement: $\mathbf{x}'$ being on the data manifold. The optimization is highly non-linear due to the compounded non-linearity of the classification boundary and the data manifold. The former dictates that any greedy search (e.g. gradient-based) near the boundary will suffer from local minima; while the latter means that not all perturbation directions result in equal visual quality (in-manifold perturbations tend to be better than off-manifold perturbations). Consequently, there are often conflicts between these two spaces when searching for $\mathbf{x}'$. To reconcile the conflicts, we propose a new simple yet effective method called *guided manifold walk* (GMW) which can compute $\mathbf{x}'$ that is close to $\mathbf{x}$ and also on the data manifold.

We exhaustively evaluate BASAR on state-of-the-art classifiers on multiple datasets in both untargeted and targeted attack tasks. The results show that not only is BASAR successful across models and datasets, it can also find on-manifold adversarial samples, in contrast to the common assumption that adversarial samples mainly exist off-manifold [7]. On par with recent work that also found on-manifold samples in images [7], we show, for the first time, the existence and commonality of such samples in skeletal motions. We also comprehensively compare BASAR with other methods, showing the superiority of BASAR by large margins. In addition, since the perturbation magnitude alone is not enough to evaluate the attack quality, we propose a new protocol for perceptual study and conduct harsh perceptual evaluation on the naturalness, deceitfulness, and indistinguishability of the attack. The perceptual results show that on-manifold adversarial examples seem more natural and realistic than regular adversarial examples. Further motivated by this observation, we recognize that on/off-manifold adversarial examples have different distributions, which forms a new more *fine-grained* description of the adversarial sample distribution. Consequently, we propose a new adversarial training approach called mixed manifold-based adversarial training (MMAT) to explore the interactions between on/off-manifold adversarial samples and clean samples during the adversarial training. We show that a proper mixture of adversarial samples with clean samples can simultaneously improve the accuracy and robustness, as opposed to the common assumption that there is always a trade-off between them [8]. Overall, the philosophy behind MMAT is general and can be potentially employed on other data/tasks, e.g. images.

This paper is an extension of our prior research [6,9]. The improvements and extensions include: (1) a new adversarial training method for skeleton-based HAR and detailed defense evaluation, (2) new attack experiments in more competitive classifiers and datasets, (3) new experiments integrating manifold projection with other attacks, (4) new literature review on Adversarial Defense, (5) new discussions about limitations, (6) additional details of mathematical deduction, implementation, and performance and (7) details for perceptual studies.

2. Related work

2.1. Skeleton-based activity recognition

Early HAR research focuses on useful hand-crafted features. In the era of deep learning, features are automatically learned. Motions can be treated as time series of joint coordinates and modeled by Recurrent Neural Networks [10]. Motions can also be converted into pseudo-images and learned with Convolutional Neural Networks [11]. Graph Convolutional Networks (GCN) recently achieve state-of-the-art performance, by considering the skeleton as a graph (joints as the nodes and bones as edges) [12–15]. Our work is complementary to HAR, by demonstrating their vulnerability to adversarial attacks and suggesting potential improvements. We extensively evaluated BASAR on state-of-the-art methods, showing that even the very recent methods with remarkable successes are still vulnerable to adversarial attacks.

2.2. Adversarial attack

Since [1], an increasing number of adversarial attack methods have been proposed in different tasks [2], such as in medical image [16] and person re-identification [17]. Goodfellow et al. [18] generate adversarial examples using the gradient of the model. Most of them consider the white-box setting, where the model is accessible to the attacker. Apart from common computer vision tasks such as classification [18–20], adversarial attacks on general time series [3] and HAR [6] have attracted attention recently. Specifically on skeleton-based HAR, an adapted version of [19] is proposed in [4] to attack skeletal motions. Wang et al. [6] introduced a novel perceptual loss to achieve effective and imperceptible attack. Despite existing successes, current methods are based on the full access to the attacked models, and therefore are not very applicable in real-world scenarios since the details of classifiers are not usually exposed to the attacker.

The difficulties of white-box attack in the real-world motivate the black-box attack, where attackers cannot access the full information of the attacked model. A simple approach is transfer-based attack, which generates adversarial samples from one surrogate model via white-box attack [6]. Existing black-box methods on skeletal motions [4,6] all rely on such a method, and cannot guarantee success due to the heavy dependence on the surrogate model [6]. In a truly black-box setting, only the final class labels (hard-labels) can be used, such setting is also called hard-label attack. Brendel et al. [21] perform the first hard-label attack by a random walk along the decision boundary. The Rays attack [22] employs a discrete search algorithm to reduce unnecessary searches. However, existing hard-label attacks do not explicitly model the data manifold, and are hence incapable of considering the visual imperceptibility of the attack if they are adapted to attack skeletal motions.

2.3. Adversarial training

The original idea of adversarial training (AT) [1] is to train classifiers with a mixture of adversarial samples and clean data, to defend against adversarial attacks. Goodfellow et al. [18] further extended the approach by using an attacker to generate adversarial examples during AT. Madry et al. [23] later redefined AT using robust optimization. Despite the significant progresses in AT [2,24], existing methods all compromise the accuracy to different extents. More importantly, the defense for skeleton-based HAR has still been largely under-explored. Early research [8] postulates that the trade-off between adversarial robustness and accuracy may be inherent. However, some recent works have proven that the trade-off can be mitigated or even theoretically eliminated. A series of works [25,26] have demonstrated that using extra (synthesis) data can mitigate such a trade-off. Stutz et al. [27] showed the existence of on-manifold adversarial samples, and reckon
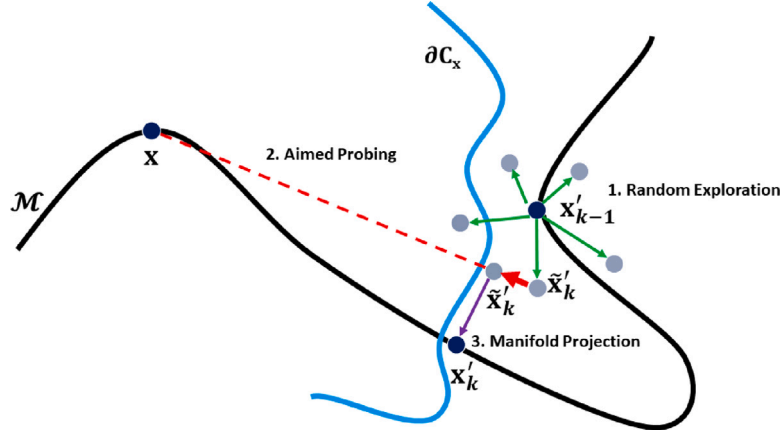
Fig. 1. An abstract 2D illustration of BASAR. $\mathbf{x}$ is the attacked motion. $\mathbf{x}'_k$ is the ideal adversarial sample in iteration k. $\mathcal{M}$ (black line) is the natural pose manifold and $\partial C_{\mathbf{x}}$ (blue line) is the class boundary of $c_{\mathbf{x}}$. $\mathbf{x}'_{k-1}$ is the result of last iteration. $\tilde{\mathbf{x}}'_k$ is the intermediate result of the current iteration. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

that on-manifold robustness is essentially related to the model generalization. Yang et al. [28] found if different classes are at least $2r$ apart, then there exists an ideal classifier which can defend against any attacks bounded by r without compromising the accuracy. Pang et al. [29] attribute the trade-off to the improper definition of robustness, hence they substitute inductive bias of local invariance with local equivariance to redefine the robust error. Very recently, adversarial defenses for video modality have just been attempted [30,31]. In this paper, we further extend [6,9] to propose a new on-manifold adversarial training for skeleton-based HAR. The results show the proposed defense can potentially truly eliminate the trade-off between robustness and accuracy for skeletal motions.

3. Methodology

We denote a motion with n poses as $\mathbf{x} = \{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$, where each pose $\mathbf{x}^t = \{q_1, \ldots, q_m\}$ includes m Dofs (joint positions or angles). A trained activity classifier G maps a motion to a probabilistic distribution over classes, $G: \mathbf{x} \to \mathbb{R}^C$ where C is the total number of action classes. The class label $c_{\mathbf{x}}$ then can be derived e.g. via *softmax*. An adversarial sample $\mathbf{x}'$ corresponding to $\mathbf{x}$ can be found via [20]:

$$\text{minimize} \quad L(\mathbf{x}, \mathbf{x}') \quad \text{s.t.} \quad c_{\mathbf{x}'} = c \text{ and } \mathbf{x}' \in [0, 1]^{m \times n} \tag{1}$$

where L is the Euclidean distance. c is the targeted class. Note that the constraint can also be replaced by $c_{\mathbf{x}'} \neq c_{\mathbf{x}}$ for untargeted attack. However, simply applying this attack to skeletal motions is not sufficient because it only restricts the adversarial sample $\mathbf{x}'$ in a hyper-cube $[0, 1]^{m \times n}$. Given that human poses lie in a natural pose manifold $\mathcal{M}$, $\mathbf{x}'$ can easily contain off-manifold poses which are unnatural/implausible and easily perceptible. We therefore add another constraint $\mathbf{x}' \in \mathcal{M}$:

$$\text{minimize} \quad L(\mathbf{x}, \mathbf{x}') \quad \text{s.t.} \quad \mathbf{x}' \in [0, 1]^{m \times n}, \mathbf{x}' \in \mathcal{M}$$

$$c_{\mathbf{x}'} = c \text{ (targeted) or } c_{\mathbf{x}'} \neq c_{\mathbf{x}} \text{ (untargeted).} \tag{2}$$

In practice, we find that $\mathbf{x}' \in [0, 1]^{m \times n}$ is less restrictive than other constraints and always satisfied. The optimization is highly nonlinear and cannot be solved analytically. It thus requires a numerical solution.

3.1. Guided manifold walk

We propose a new method called *Guided Manifold Walk* (GMW) to solve Eq. (2). For simplicity, we start with an abstract 2D illustration of GMW on $\mathbf{x}$ shown in Fig. 1. $\mathbf{x}'_k$ is the ideal adversarial sample which is on-manifold and close to $\mathbf{x}$ in the kth iteration. Given the non-linearity of the classification boundary and the data manifold, BASAR aims to exploit the properties of both simultaneously. The GMW is an iterative

approach where two major steps are alternatively conducted. One step is to find an adversarial sample that is close to $\mathbf{x}$ and the other one is to find the closest sample on the data manifold from an arbitrary off-manifold position. Since the former mainly considers the classification boundary, we design two sub-routines: *random exploration* and *aimed probing*. Similar sampling strategies have been attempted in attacking images [21]. We extend them to motions by treating a whole motion as an $n \times m$ image, with joint weighting and automatically thresholded attacks. Random exploration is to explore the vicinity of the current adversarial sample to find a random sample (step 1 in Fig. 1). Aimed probing is to find a sample in proximity to $\partial C_{\mathbf{x}}$ and is closer to $\mathbf{x}$ (step 2 in Fig. 1). The details are given in Sections 3.2 and 3.3. Finally, we design a sub-routine: *manifold projection* which is to project an off-manifold sample $\tilde{\mathbf{x}}'_k$ onto $\mathcal{M}$ to obtain $\mathbf{x}'_k$ (step 3 in Fig. 1). This is one key element of our approach in bringing the motion manifold into adversarial attack. The algorithm overview is given in Algorithm 1, where λ and β are hyper-parameters and l is a distance function. Next, we give details of all sub-routines.

3.2. Random exploration

Random exploration is to explore in proximity to the classification boundary, by making a small step towards a random direction:

$$\tilde{\mathbf{x}} = \mathbf{x}' + \mathbf{W}\Delta, \text{where } \Delta = \mathbf{R} - (\mathbf{R}^T \mathbf{d})\mathbf{d}, \ \mathbf{d} = \frac{\mathbf{x} - \mathbf{x}'}{\|\mathbf{x} - \mathbf{x}'\|},$$

$$\mathbf{R} = \lambda \frac{\mathbf{r}}{\|\mathbf{r}\|} \|\mathbf{x} - \mathbf{x}'\|, \ r \in N(0, \mathbf{I}), \tag{3}$$

where $\tilde{\mathbf{x}}$ is the new perturbed sample, $\mathbf{x}$ and $\mathbf{x}'$ are the attacked motion and current adversarial sample. The perturbation on $\mathbf{x}'$, Δ, is weighted by $\mathbf{W}$ - a diagonal matrix with joint weights. This is based on the experiments and perceptual study which suggest that equal perturbations on different joints are not equally effective and imperceptible, e.g. perturbations on the spinal joints cause larger visual distortion but are less effective in attacks. We therefore weight them differently. Δ controls the direction and magnitude of the perturbation, and depends on two variables $\mathbf{R}$ and $\mathbf{d}$. $\mathbf{d}$ is the directional vector from $\mathbf{x}'$ to $\mathbf{x}$. $\mathbf{R}$ is a random directional vector sampled from a Normal distribution $N(0, \mathbf{I})$ where $\mathbf{I}$ is an identity matrix, $\mathbf{I} \in R^{z \times z}$, $z = mn$, m is the number of Dofs in one frame and n is total frame number. This directional vector is scaled by $\|\mathbf{x} - \mathbf{x}'\|$ and λ.

3.3. Aimed probing

Aimed probing is straightforward, aiming to find a new adversarial sample between the perturbed motion and the original, so that the

Algorithm 1: Overview of the GMW

Input: $\mathbf{x}$: attacked motion;
$\widetilde{\mathbf{x}}_0$: random sample, where
$\quad c_{\widetilde{\mathbf{x}}_0} = c$ (targeted) or $c_{\widetilde{\mathbf{x}}_0} \neq c_{\mathbf{x}}$ (untargeted);
K: maximum number of iterations; ϵ: perturbation threshold; λ, β_1 and β_2:
$\quad$ hyper-parameters;
Initialization: $\mathbf{x}'_0 = AimedProbing(\widetilde{\mathbf{x}}_0, \mathbf{x}, \beta_1)$, so that $\mathbf{x}'_0$ is adversarial and $\beta_1 =$
$\quad$ update$(\beta_1, \mathbf{x}'_0)$;
for $k = 1$ **to** K **do**
$\quad$ $\widetilde{\mathbf{x}}_k = RandomExploration(\mathbf{x}'_{k-1}, \mathbf{x}, \lambda)$;
$\quad$ **while** $\widetilde{\mathbf{x}}_k$ *is not adversarial and* $\lambda \geq 10^{-10}$ **do**
$\quad\quad$ $\lambda = $ update$(\lambda, \widetilde{\mathbf{x}}_k)$;
$\quad\quad$ $\widetilde{\mathbf{x}}_k = RandomExploration(\mathbf{x}'_{k-1}, \mathbf{x}, \lambda)$;
$\quad$ **end**
$\quad$ **if** $\lambda \geq 10^{-10}$ **then** $\mathbf{x}'_k = \widetilde{\mathbf{x}}_k$; $\lambda = $ update$(\lambda, \widetilde{\mathbf{x}}_k)$;
$\quad$ **else** $\mathbf{x}'_k = \mathbf{x}'_{k-1}$; break;
$\quad$ $\widetilde{\mathbf{x}}_k = AimedProbing(\mathbf{x}'_k, \mathbf{x}, \beta_1)$;
$\quad$ **while** $\widetilde{\mathbf{x}}_k$ *is not adversarial and* $\beta_1 \geq 10^{-10}$ **do**
$\quad\quad$ $\beta_1 = $ update$(\beta_1, \widetilde{\mathbf{x}}_k)$;
$\quad\quad$ $\widetilde{\mathbf{x}}_k = AimedProbing(\mathbf{x}'_k, \mathbf{x}, \beta_1)$;
$\quad$ **end**
$\quad$ **if** $\beta_1 \geq 10^{-10}$ **then** $\mathbf{x}'_k = \widetilde{\mathbf{x}}_k$; $\beta_1 = $ update$(\beta_1, \widetilde{\mathbf{x}}_k)$;
$\quad$ **else** break;
$\quad$ $\widetilde{\mathbf{x}}_k = ManifoldProjection(\mathbf{x}'_k, \mathbf{x})$; $\hat{\mathbf{x}}_k = \widetilde{\mathbf{x}}_k$;
$\quad$ **while** $\widetilde{\mathbf{x}}_k$ *is not adversarial and* $\beta_2 \geq 10^{-10}$ **do**
$\quad\quad$ $\beta_2 = $ update$(\beta_2, \hat{\mathbf{x}}_k)$;
$\quad\quad$ $\widetilde{\mathbf{x}}_k = AimedProbing(\mathbf{x}'_k, \hat{\mathbf{x}}_k, \beta_2)$;
$\quad$ **end**
$\quad$ **if** $\beta_2 \geq 10^{-10}$ **then** $\mathbf{x}'_k = \widetilde{\mathbf{x}}_k$; $\beta_2 = $ update$(\beta_2, \widetilde{\mathbf{x}}_k)$;
$\quad$ **else** break;
$\quad$ **if** $l(\mathbf{x}'_k, \mathbf{x}) < \epsilon$ **then** break;
end
return $\mathbf{x}'_k$;

new sample is closer to the attacked motion and remains adversarial: $\widetilde{\mathbf{x}} = \mathbf{x}' + \beta(\mathbf{x} - \mathbf{x}')$, where β is a forward step size. Similar to λ, β is decreased to conduct the aimed probing again if $\widetilde{\mathbf{x}}$ is not adversarial; otherwise, we increase β, then enter the next sub-routine.

3.4. Manifold projection

After aimed probing and random exploration, the perturbed motion $\widetilde{\mathbf{x}}$ is often off the manifold, resulting in implausible and unnatural poses. We thus project them back to the manifold. The natural pose manifold can be regarded as poses that do not violate bone lengths or joint limits, i.e. they are realizable by humans. Further, a motion is regarded as on-manifold if all its poses are on-manifold. The motion manifold can be obtained in two ways: explicit modeling or implicit learning. Using implicit learning would require to train a data-driven model then use it for projection, breaking BASAR into a two-step system. Therefore we employ explicit modeling. Specifically, we replace the constraint $\mathbf{x}' \in \mathcal{M}$ in Eq. (2) with hard constraints on bone lengths and joint limits. We also constrain the dynamics of $\mathbf{x}'$ to be similar to the original motion $\mathbf{x}$:

$$\min_{\mathbf{x}'} \quad L(\widetilde{\mathbf{x}}, \mathbf{x}') + wL(\ddot{\mathbf{x}}, \ddot{\mathbf{x}}')$$
s.t. $\quad B'_i = B_i$ and $\theta_j^{\min} \leq \theta'_j \leq \theta_j^{\max}$

$C_{\mathbf{x}'} = c$ (targeted) or $C_{\mathbf{x}'} \neq C_{\mathbf{x}}$ (untargeted), $\quad\quad\quad\quad$ (4)

where $\ddot{\mathbf{x}}$ and $\ddot{\mathbf{x}}'$ are the 2nd-order derivatives of $\mathbf{x}$ and $\mathbf{x}'$, w is a weight. Matching the 2nd-order derivatives is proven to be important for visual imperceptibility in adversarial attack [6]. L is the Euclidean distance. B_i and B'_i are the ith bone's lengths of the attacked and adversarial motion respectively. When the bone lengths change from frame to frame in the original data, we impose the bone-length constraint on each frame. θ'_j is the jth joint angle in every pose of $\mathbf{x}'$ and subject to joint limits bounded by $\theta_j^{\min}$ and $\theta_j^{\max}$. Essentially, the optimization above seeks an adversarial sample that is: (1) close to the perturbed

motion $\widetilde{\mathbf{x}}$ in terms of the Euclidean distance, (2) matching the motion dynamics to the original motion $\mathbf{x}$ and (3) on the manifold.

Eq. (4) is difficult to solve, especially to satisfy both the bone length and joint limit constraints in the joint position space [5]. We therefore solve Eq. (4) in two steps. First, we solve it without any constraints by Inverse Kinematics in the joint angle space, which automatically preserves the bone lengths. Next, Eq. (4) is solved in the joint angle space:

$$\min_{\theta'} \quad L(\widetilde{\theta}, \theta') + wL(\ddot{\theta}, \ddot{\theta}') \quad \text{s.t.} \quad \theta_j^{\min} \leq \theta'_{\mathbf{j}} \leq \theta_j^{\max},$$

$C_{\mathbf{x}'} = c$ (targeted) or $C_{\mathbf{x}'} \neq C_{\mathbf{x}}$ (untargeted). $\quad\quad$ (5)

Note that the objective function in Eq. (5) is designed to match the joint angles and the joint angular acceleration. We use a primal–dual interior-point method [32] to solve Eq. (5). After solving for θ', the joint positions of the adversarial motion are computed using Forward Kinematics. Please refer to the supplementary materials for details of mathematical deduction.

3.5. Mixed on-manifold adversarial training

The assumption of adversarial training is that adversarial samples can help regulate classification boundaries to resist attacks [27]. A common adversarial training (AT) strategy [23] is:

$$\min_{\theta} \mathbb{E}_{(\mathbf{x},y) \sim D} \left[\max_{\delta \in B(\epsilon)} \mathcal{L}(\mathbf{x} + \delta, \theta, y) \right] \quad\quad\quad (6)$$

where D is the distribution over data pairs of $\mathbf{x} \in \mathbb{R}^d$ and label y. $\theta \in \mathbb{R}^p$ is the model parameters. $B(\epsilon) = \{\delta \mid \|\delta\|_\infty \leq \epsilon \}$ is the perturbation set. $\mathcal{L}$ is a loss function, e.g. cross-entropy. During training, the perturbation δ is drawn from a prior e.g. Gaussian and uniform distribution, or some adversarial attack method, and restricted within the ϵ-ball $B(\epsilon)$.

One issue in Eq. (6) is that there is an underlying assumption of the structural simplicity of the adversarial sample distribution in $B(\epsilon)$, which enables the usage of well-defined priors (e.g. Gaussians). However, we argue this assumption is overly simplified. The structure of the adversarial sample distribution can be arbitrarily complex. Consequently, when drawing perturbations from a prior, a conservative prior (e.g. Gaussians with too small variances) cannot resist attacks, while an aggressive one (e.g. Gaussians with too large variances) can be detrimental to the accuracy. On the other hand, drawing perturbations via attack leads to a more guided approximation of the adversarial sample distribution, but it also ignores the different importance across different adversarial samples. As a result, existing AT methods always need to compromise between accuracy and robustness [8].

Unlike existing methods, we explore a finer structure depicted by two distributions in $B(\epsilon)$: the distributions of on-manifold and off-manifold adversarial samples. We first assume the distribution of adversarial samples is different from the clean data [7]. Next, we further make a more *fine-grained* assumption: the distribution of on-manifold adversarial samples is different from that of the off-manifold adversarial samples. We expect the fine-grained distribution modeling to be able to eliminate the trade-off between accuracy and robustness, which remains unsolved currently. This is because on-manifold samples should be directly useful in simultaneously improving the accuracy and robustness, while the off-manifold samples are more aggressive and hence helpful in improving the robustness.

To this end, we propose a mixed manifold-based adversarial training (MMAT) which optimizes a hybrid loss consisting of a standard classification loss ($\mathcal{L}_c$), an on-manifold robustness loss ($\mathcal{L}_{on}$) and an off-manifold robustness loss ($\mathcal{L}_{off}$) term:

$$\min_{\theta} \mathbb{E}_{(\mathbf{x},y) \sim D} \left[\mu_c \mathcal{L}_c + \mu_{on} \mathcal{L}_{on} + \mu_{off} \mathcal{L}_{off} \right] \quad\quad (7)$$

where μ_c, μ_{on} and μ_{off} are weights, $\mu_c = 1 - \mu_{on} - \mu_{off}$. The losses are:

$$\mathcal{L}_c = \mathcal{L}(\mathbf{x}, \theta, y) \quad\quad\quad\quad (8)$$

$$\mathcal{L}_{on} = \mathcal{L}(\mathbf{x}'_{on}, \boldsymbol{\theta}, y) \text{ s.t. } \mathbf{x}'_{on} \in [0,1]^{m \times n}, \ \mathbf{x}'_{on} \in \mathcal{M} \tag{9}$$

$$\mathcal{L}_{off} = \max \mathcal{L}(\mathbf{x}'_{off}, \boldsymbol{\theta}, y) \text{ s.t. } \mathbf{x}'_{off} \in [0,1]^{m \times n} \tag{10}$$

Adversary Sampling: During optimization, we need to sample $\mathbf{x}'_{on}$ and $\mathbf{x}'_{off}$ as they cannot be described in any analytical form. We propose a black-box and a white-box sampling strategy: BASAR and SMART [6] with MMAT, named BASAR-MMAT and SMART-MMAT respectively. In BASAR-MMAT, $\mathbf{x}'_{on}$ and $\mathbf{x}'_{off}$ are generated by BASAR with/without manifold projection (BASAR-NoMP). In SMART-MMAT, $\mathbf{x}'_{on}$ and $\mathbf{x}'_{off}$ are generated by SMART [6] with/without perception loss.

4. Attack experiments

4.1. Settings

To evaluate the proposed method, we conduct experiments on five state-of-the-art target models: ST-GCN [12], MS-G3D [33], SGN [13], CTR-GCN [14] and FR-HEAD [15]. We choose not only the most popular benchmark datasets such as HDM05 [34] and NTU 60 [35], but also the challenging dataset UAV-Human [36], Kinetics-400 [37] and Skeletics-152 [38]. UAV-Human [36] was collected by an unmanned aerial vehicle and hence has low visual quality. Kinetics-400 [37] is a large and highly noisy dataset taken from different YouTube Videos. Similarly, Skeletics-152 is a 3D pose-annotated subset of videos sourced from much larger Kinetics-700 datasets [39]. The details of these databases are in the supplemental document. The experiments are conducted on a Xeon Silver 4216 CPU and an NVIDIA GTX 2080Ti GPU. The query time and number of queries are shown in the supplemental document.

4.2. Evaluation metrics

We employ the success rate as for evaluation. In addition, to further numerically evaluate the quality of the adversarial samples, we also define metrics between the original motion $\mathbf{x}$ and its adversarial sample $\mathbf{x}'$, including the averaged joint position deviation $l = \frac{1}{nN} \sum_{j=0}^{N} \|\mathbf{x}^{(j)} - \mathbf{x}'^{(j)}\|_2$, the averaged joint acceleration deviation $\Delta a = \frac{1}{nON} \sum_{j=0}^{N} \|\ddot{\mathbf{x}}^{(j)} - \ddot{\mathbf{x}}'^{(j)}\|_2$, the averaged joint angular acceleration deviation $\Delta \alpha = \frac{1}{nON} \sum_{j=0}^{N} \|\ddot{\boldsymbol{\theta}}^{(j)} - \ddot{\boldsymbol{\theta}}'^{(j)}\|_2$, and the averaged bone-length deviation percentage $\Delta B / B = \frac{\sum_{j=0}^{N} \sum_{i=0}^{T} ((B_i^{(j)} - B_i'^{(j)})/B_i^{(j)})}{T N}$, where N is the number of adversarial samples. O and T are the total number of joints and bones in a skeleton. n is the number of poses in a motion. We also investigate the percentage of on-manifold (OM) adversarial motions. An attack sample is regarded as on-manifold if all its poses respect the bone-length and joint limit constraints. Finally, since Kinetics-400, UAV-Human and Skeletics-152 have missing joints, it is impossible to attack it in the joint angle space. So we only attack it in the joint position space. Consequently, $\Delta \alpha$ and OM cannot be computed on Kinetics-400, UAV-Human and Skeletics-152.

4.3. Attack evaluation

To initialize for untargeted attack, we randomly sample a motion $\mathbf{x}'$ for a target motion $\mathbf{x}$ where $c_{\mathbf{x}'} \neq c_{\mathbf{x}}$. For Kinetics-400, UAV-Human and Skeletics-152, the maximum number of iterations is 1000, and it is set to 500 and 2000 for HDM05 and NTU 60 respectively. The results are shown in Table 1. Note that BASAR achieves 100% success in all tasks. Here we also conduct ablation studies (MP/No MP) to show the effects of the manifold projection. First, the universal successes across all datasets and models demonstrate the effectiveness of BASAR. The manifold projection directly affects the OM results. BASAR can generate as high as 99.55% on-manifold adversarial samples. As shown in the perceptual study later, the on-manifold samples are very hard to be distinguished from the original motions even under very harsh visual comparisons.

For targeted attack, the maximum iterations are set to 1000, 2000, 3000 on HDM05, Kinetics-400 and NTU 60. To initiate a targeted attack on $\mathbf{x}$, we randomly select a $\mathbf{x}'$ where $C_{\mathbf{x}'} = c$ and c is the targeted class. The results are shown in Table 1. All attacks achieve 100% success. The targeted attack is more challenging than the untargeted attack [6], because the randomly selected label often has completely different semantic meanings from the original one. Attacking an 'eating' motion to 'drinking' is much easier than to 'running'. This is why the targeted attack, in general, has worse results than untargeted attack under every metric. Even under such harsh settings, BASAR can still produce as high as 56.98% on-manifold adversarial samples. The performance variation across models is consistent with the untargeted attack.

Attack on latest classifiers and datasets. CTR-GCN [14] and FR-HEAD [15] are the recently proposed classifiers with more robust skeleton representations. We hence investigate the effectiveness of BASAR on the two challenging target models. As shown in Table 2, BASAR can still generate manifold adversarial samples with a high probability even when the target models have robust representations. We also test BASAR on the latest wild human motion dataset UAV-Human [36] and Skeletics-152 [38], and the results are reported in Table 3. The data quality of UAV-Human and Skeletics-152 is similar to Kinetics-400 and so is the attack performance. Overall, BASAR with manifold projection can improve the attack quality via reducing the l, Δa and $\Delta B/B$.

4.4. Perceptual studies

Numerical evaluation alone is not sufficient to evaluate the imperceptibility of adversarial attack on skeletal motions, because they cannot accurately indicate whether the attack is perceptible to humans [6]. We, therefore, conduct rigorous perceptual studies to evaluate the imperceptibility of BASAR. Therefore, we design a new perceptual study protocol that includes three perception metrics: Deceitfulness, Naturalness and Indistinguishability. Deceitfulness is to test whether BASAR visually changes the semantics of the motion. Naturalness aims to test whether on-manifold adversarial samples look more natural than off-manifold adversarial samples. Indistinguishability is the strictest test to see whether adversarial samples by BASAR can survive a side-by-side scrutiny. The details of the perceptual study is reported in **Supplemental Document**.

The success rate of *Deceitfulness* is 79.64%. Next, 85% of the on-manifold adersarial samples look more natural than off-manifold samples. This is understandable as manifold projection not only makes sure the poses are on the manifold, but also enforces the similarity of the dynamics between the attacked and original motion. Finally, the results of *Indistinguishability* are 89.90% on average. BASAR even outperforms the white-box attack (80.83%) in [6]. BASAR successfully fools the users under the strictest side-by-side scrutiny. The complete perceptual evaluation is reported in **Supplemental Document**

4.5. Comparison

Since BASAR is the very first black-box adversarial attack method on skeletal motions, there is no baseline for comparison. So we employ methods that are closet to our approach as baselines. Although SMART [6] is designed for white-box attack, it can also be used as transfer-based attack via requiring surrogate models. We hence employ it as a baseline and choose HRNN [40] and 2SA-GCN [41] as the surrogate models. The second baseline is MTS [3] which is a black-box method but only on general time-series. It is the most similar method to BASAR but does not model the data manifold. Another baseline is BA [21], a decision-boundary based attack designed for images. We choose HDM05, NTU 60 and Skeletics 152 for comparisons. Since MTS is not designed for untargeted attack, we only compare BASAR with it on the targeted attack.

Table 1

Untargeted attack (left) and targeted attack (right) on HDM05 (top), NTU (middle) and Kinetics-400 (bottom). All attacks have a 100% success rate. l: averaged joint position deviation; Δa: averaged joint acceleration deviation; $\Delta\alpha$: averaged joint angular acceleration deviation; $\Delta B/B$: averaged bone-length deviation percentage; on-manifold sample percentage (OM). MP means Manifold Projection.

Models		Untargeted attack					Targeted attack				
		$l \downarrow$	$\Delta a \downarrow$	$\Delta\alpha \downarrow$	$\Delta B/B \downarrow$	OM $\uparrow$	$l \downarrow$	$\Delta a \downarrow$	$\Delta\alpha \downarrow$	$\Delta B/B \downarrow$	OM $\uparrow$
ST-GCN	MP	0.13	0.05	0.11	0.00%	99.55%	4.97	0.10	0.65	3.44%	56.98%
	No MP	0.10	0.04	0.34	0.66%	0.00%	6.25	0.09	0.92	5.85%	0.00%
MS-G3D	MP	0.76	0.12	0.49	1.78%	0.13%	4.34	0.12	0.71	4.51%	1.64%
	No MP	0.70	0.09	0.82	1.81%	0.00%	4.35	0.11	1.01	5.08%	0.00%
SGN	MP	11.53	1.92	6.70	9.60%	60.52%	16.31	1.28	6.97	12.29%	20.96%
	No MP	7.93	2.00	14.36	39.64%	0.00%	16.13	1.63	13.28	29.86%	0.00%
ST-GCN	MP	0.08	0.02	0.07	4.82%	4.68%	0.37	0.03	0.25	9.73%	0.63%
	No MP	0.10	0.02	0.09	5.57%	1.82%	0.38	0.04	0.16	11.55%	0.16%
MS-G3D	MP	0.08	0.03	0.12	8.14%	0.86%	0.36	0.05	0.24	15.43%	0.00%
	No MP	0.12	0.03	0.17	10.02%	0.57%	0.40	0.06	0.27	17.72%	0.00%
SGN	MP	0.28	0.08	0.21	11.11%	28.95%	1.28	0.09	0.38	28.24%	2.63%
	No MP	0.30	0.10	0.42	28.00%	4.55%	1.35	0.10	0.53	39.43%	0.00%
ST-GCN	MP	0.05	0.0057	n/a	2.54%	n/a	0.63	0.03	n/a	29.10%	n/a
	No MP	0.07	0.0062	n/a	3.53%	n/a	0.67	0.03	n/a	31.48%	n/a
MS-G3D	MP	0.10	0.011	n/a	5.16%	n/a	0.56	0.05	n/a	27.26%	n/a
	No MP	0.10	0.012	n/a	5.69%	n/a	0.57	0.07	n/a	28.35%	n/a
SGN	MP	0.12	0.020	n/a	4.23%	n/a	1.51	0.18	n/a	68.45%	n/a
	No MP	0.13	0.022	n/a	6.93%	n/a	1.54	0.19	n/a	72.09%	n/a

Table 2

(Untargeted) Attack performance on latest classifiers. All attacks have a 100% success rate.

Models		HDM05					NTU 60				
		$l \downarrow$	$\Delta a \downarrow$	$\Delta\alpha \downarrow$	$\Delta B/B \downarrow$	OM $\uparrow$	$l \downarrow$	$\Delta a \downarrow$	$\Delta\alpha \downarrow$	$\Delta B/B \downarrow$	OM $\uparrow$
CTR-GCN	MP	0.67	0.14	0.31	0.80%	45.0%	0.05	0.02	0.03	6.50%	2.2%
	No MP	0.63	0.13	1.00	2.18%	1.4%	0.07	0.02	0.04	7.14%	0.8%
FR-HEAD	MP	0.15	0.06	0.07	0%	93.2%	0.04	0.013	0.06	4.13%	10.8%
	No MP	0.16	0.05	0.42	0.76%	5.9%	0.06	0.014	0.07	5.14%	7.9%

Table 3

(Untargeted) Attack performance on the wild human motion dataset with ST-GCN. All attacks have a 100% success rate.

ST-GCN	UAV			Skeletics-152		
	$l \downarrow$	$\Delta a \downarrow$	$\Delta B/B \downarrow$	$l \downarrow$	$\Delta a \downarrow$	$\Delta B/B \downarrow$
MP	17.68	9.30	6.81%	0.095	0.019	3.73%
NO MP	21.83	11.41	8.05%	0.098	0.019	3.88%

Table 4 lists the success rates of all methods. BASAR performs the best and often by big margins. In the targeted attack, the highest attack success rate among the baseline methods is merely 30.3% on HDM05, 12.9% on NTU and 4.7% on Skeletics-152 while BASAR achieves 100%. In the untargeted attack, the baseline methods achieve higher performances but still worse than BASAR. SMART achieves as high as 99.33% on NTU/MS-G3D. However, its performance is not reliable as it highly depends on the chosen surrogate model, which is consistent with [6]. In addition, we further look into the results and find that SMART's results are inconsistent. When the attack is transferred, the result labels are often different from the labels obtained during the attack.

We find that BA can also achieve 100% success. However, BA is designed to attack image data and does not consider the data manifold. We therefore compare detailed metrics and show the results in Table 5. BA is in general worse than BASAR under every metric. The worst is the bone-length constraint violation. Visually, the skeletal structure cannot be observed at all. This happens for both the untargeted and the targeted attack across all datasets and models. This is understandable because BA does not consider the data manifold, while BASAR assumes that in-manifold perturbations provide better visual quality. One qualitative comparison with BA can be found in Fig. 2.

4.5.1. Effectiveness of manifold projection

The manifold projection is a general operation which can theoretically work with other attackers. To verify this, we adapt Rays [22], a state-of-the-art decision-based attack for images, to attack motions by treating a motion sample as an image. Based on our experiments, Rays fails in targeted attack even with manifold projection, which is not surprising as it is not designed for attacking motions. Therefore, we only report the untargeted attack results. The model queries are the same as BASAR. As shown in Table 6, Rays-MP with manifold projection can always improve the attack success rate and the attack quality via reducing the l, Δa, $\Delta\alpha$ and $\Delta B/B$ metrics. Moreover, Rays-MP can generate more natural adversarial motions.

5. Defense experiments

5.1. Experiment setup

We evaluate MMAT on HDM05 [34] using ST-GCN [12], MS-G3D [33] and SGN [13]. NTU 60 [35] and Kinetics-400 [37] are excluded for their extensive noises making it difficult to evaluate the effects of on-manifold samples in AT. We follow the original settings [12,13,33] to train these networks. For BASAR-MMAT, we regard adversarial examples generated by BASAR as data augmentation. For SMART-MMAT, we use SMART-50 (SMART with 50 iterations) for training. After adversarial training, we attack the trained model with BASAR-NoMP with 500 iterations, as it generates more violent attacks than BASAR. We also employ SMART-200 [6] and CIASA-200 [4] to test the classifier robustness under white-box attacks. Since there are three weights in our adversarial training loss, we conduct an ablation study to identify the optimal weights in different settings in **Supplemental Document**.

Table 4
Attack success rate comparison with baseline methods. S-152 means the Skeletics-152 dataset.

Models	Attacked method	Untargeted attack			Targeted attack		
		HDM05	NTU	S-152	HDM05	NTU	S-152
ST-GCN	BASAR	100%	100%	100%	100%	100%	100%
	MTS	n/a	n/a	n/a	3.3%	12.0%	4.7%
	SMART(HRNN)	66.9%	89.3%	33.6%	3.2%	2.3%	2.2%
	SMART(2SAGCN)	86.1%	12.9%	14.1%	2.3%	0.2%	1.1%
MS-G3D	BASAR	100%	100%	100%	100%	100%	100%
	MTS	n/a	n/a	n/a	2.2%	12.9%	3.6%
	SMART(HRNN)	86.9%	99.3%	51.1%	30.3%	1.2%	1.1%
	SMART(2SAGCN)	88.7%	3.1%	2.3%	2.5%	0.00%	0.6%
SGN	BASAR	100%	100%	100%	100%	100%	100%
	MTS	n/a	n/a	n/a	2.91%	0.00%	1.4%
	SMART(HRNN)	89.25%	98.25%	20.3%	29.69%	1.42%	2.1%
	SMART(2SAGCN)	0.41%	97.75%	10.6%	3.28%	1.83%	1.3%

Table 5
Boundary Attack (BA) on HDM05 (left), NTU (right). UA/TA refers Untargeted/Targeted attack.

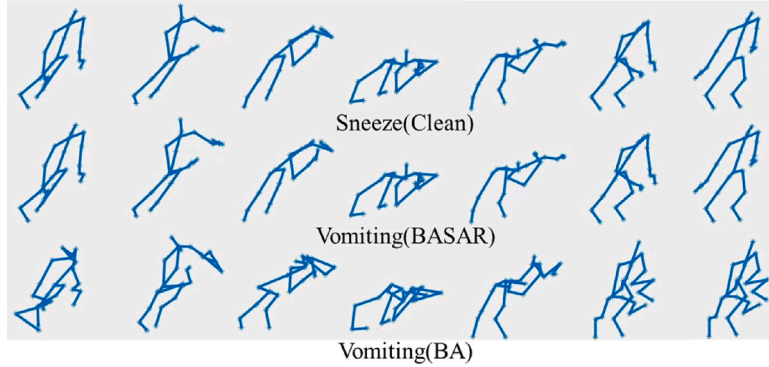| Models		$l\downarrow$	$\Delta a\downarrow$	$\Delta \alpha\downarrow$	$\Delta B/B\downarrow$	OM $\uparrow$	$l\downarrow$	$\Delta a\downarrow$	$\Delta \alpha\downarrow$	$\Delta B/B\downarrow$	OM $\uparrow$
ST-GCN	UA	1.44	0.65	4.74	10.60%	0.00%	1.04	0.47	1.97	235.10%	0.00%
	TA	8.83	0.17	1.60	8.56%	0.00%	1.04	0.47	1.97	235.10%	0.00%
MS-G3D	UA	1.17	0.36	2.81	6.00%	0.00%	1.24	1.73	2.38	911.7%	0.00%
	TA	7.93	0.10	1.07	7.49%	0.00%	0.27	0.07	0.34	25.72%	0.00%
SGN	UA	13.35	3.45	21.96	75.11%	0.00%	0.22	0.28	1.26	125.57%	3.60%
	TA	15.40	1.52	11.77	29.45%	0.00%	0.42	0.15	0.66	65.31%	0.17%

Fig. 2. The visual comparison with BA. The first row is the clean motion labeled as 'Squeeze'. The second row is the adversarial motion generated by BASAR and misclassified as 'Vomiting'. The third row is the adversarial motion generated by BA and misclassified as 'Vomiting'.

Table 6
The results of Rays and Rays with manifold projection(Rays-MP) on HDM05 (left) and NTU 60 (right). l_∞ means the l_∞ norm distance between adversarial examples and original examples. SR means attack success rate. We show the best performance with bold.

Models	Attacks	Numerical evaluation					$l_\infty(0.5)$	Numerical evaluation					$l_\infty(0.05)$
		$l\downarrow$	$\Delta a\downarrow$	$\Delta \alpha\downarrow$	$\Delta B/B\downarrow$	OM $\uparrow$		$l\downarrow$	$\Delta a\downarrow$	$\Delta \alpha\downarrow$	$\Delta B/B\downarrow$	OM $\uparrow$	
ST-GCN	Rays	0.08	0.015	0.19	0.02%	27.2%	100%	0.061	0.028	0.019	4.6%	2.2%	99.9%
	Rays-MP	**0.07**	**0.013**	**0.13**	**0.01%**	**56.7%**	100%	**0.056**	**0.026**	**0.018**	**4.2%**	**4.8%**	**100%**
MS-G3D	Rays	1.67	0.07	0.28	0.11%	0%	95.2%	0.069	0.0154	0.050	2.86%	1.2%	98.8%
	Rays-MP	**1.57**	**0.06**	**0.25**	**0.07%**	**3.1%**	**96.9%**	**0.039**	**0.0146**	0.047	**2.61%**	**3.8%**	**100%**
SGN	Rays	1.35	0.17	0.2	0.1%	60.8%	73.1%	0.19	0.003	0.0006	0.32%	95.9%	95.9%
	Rays-MP	**0.89**	**0.06**	**0.1**	**0.0%**	**94.0%**	**96.0%**	**0.07**	**0.0003**	**0.0001**	**0.27%**	**99.3%**	**98.5%**

5.2. Robustness evaluation

We employ TRADES [8] and MART [24] as baselines, which are the state-of-the-art AT methods on images. Zheng et al. [5] use an adapted random smoothing (RS) [42] approach for defending skeleton-based HAR, so we use it as one baseline. The results are shown in Table 7. First, both BASAR-MMAT and SMART-MMAT can improve the robustness (Table 7) and SMART-MMAT has slightly better overall performance. This is understandable since SMART-MMAT is white-box and computes more aggressive adversaries. However, BASAR-MMAT can achieve better accuracy than standard training and SMART-MMAT, which show BASAR-MMAT can potentially eliminate the accuracy-robustness trade-off. Next, SMART-MMAT is apparently more robust than RS and MART, and outperforms TRADES under most attack scenarios. More importantly, our method not only improves the robustness but also not compromise standard accuracy. The MMAT accuracy is
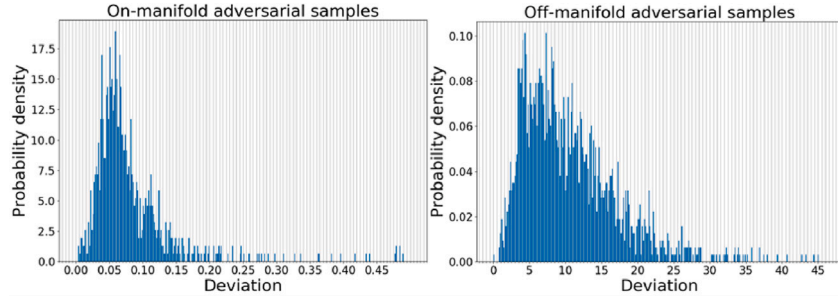
Fig. 3. Deviation distributions of on/off-manifold adversarial samples (attacking ST-GCN) on HDM05 dataset.

Table 7
Comparing robustness with different defense methods. Top: ST-GCN, Middle: MS-G3D, Bottom: SGN. Acc is the classification accuracy, i.e. average accuracy on clean data. ST means standard training. MMAT means SMART-MMAT.

Models	Methods	BASAR-NoMP			SMART		CIASA		Acc
		$l \uparrow$	$\Delta a \uparrow$	$\Delta B/B \uparrow$	@50	@200	@50	@200	
ST-GCN	ST	0.1	0.04	0.66%	3.76%	0.58%	1.69%	0.58%	87.2%
	BASAR-MMAT	2.07	0.67	10.91%	47.54%	92.63%	48.75%	93.75%	**91.2%**
	RS	2.80	1.16	19.72%	64.57%	14.33%	62.55%	11.71%	90.4%
	MART	**6.26**	2.14	39.27%	64.84%	52.50%	63.62%	49.33%	85.2%
	TRADES	5.25	1.72	30.68%	83.21%	**65.83%**	81.78%	**63.54%**	85.8%
	SMART-MMAT	5.98	**2.56**	**47.98%**	**85.68%**	62.33%	**85.31%**	60.91%	**91.0%**
MS-G3D	ST	0.70	0.09	1.81%	6.59%	1.47%	6.10%	4.19%	94.4%
	BASAR-MMAT	4.53	0.86	15.86%	77.23%	92.41%	70.76%	91.13%	**95.9%**
	RS	6.90	2.59	48.08%	90.35%	39.56%	90.57%	57.44%	93.4%
	MART	8.71	3.70	77.26%	88.09%	81.84%	87.70%	80.27%	90.6%
	TRADES	**10.43**	**4.19**	**87.34%**	88.83%	84.33%	88.63%	82.90%	90.0%
	SMART-MMAT	10.30	3.97	79.07%	**91.69%**	**84.79%**	**91.71%**	**85.06%**	93.8%
SGN	ST	7.93	2.00	39.64%	63.29%	22.56%	69.15%	22.56%	94.1%
	RS	12.87	3.07	64.18%	79.14%	32.78%	76.71%	24.76%	93.2%
	MART	8.46	2.04	40.46%	84.77%	31.25%	83.79%	27.92%	92.1%
	TRADES	8.46	2.04	40.44%	84.18%	36.53%	82.62%	25.08%	91.2%
	MMAT	**13.06**	**3.46**	**65.69%**	**87.36%**	**48.44%**	**87.00%**	**45.15%**	**93.9%**
FR-HEAD	ST	0.16	0.05	0.76%	19.0%	4.9%	14.3%	4.2%	93.4%
	RS	8.2	3.1	60.6%	88.1%	71.8%	88.2%	71.2%	92.3%
	MART	9.3	**3.3**	**67.8%**	88.7%	**76.6%**	88.1%	76%	93.0%
	TRADES	**9.4**	3.2	62.3%	89.4%	76.3%	88.7%	75.4%	93.0%
	MMAT	8.8	**3.3**	61.3%	**89.6%**	75.9%	**89.1%**	**76.3%**	93.2%

Table 8
Adaptive attack to evaluate the obfuscated gradient effect. (±xx%) means the robustness difference with SMART-200.

SMART-MMAT	ST-GCN	MS-G3D	SGN
HDM05	46.41% (−0.4%)	83.92% (−0.9%)	46.41% (−2.0%)

within a small range(+3.76%/−0.59%) from that of standard training, in contrast to the noticeable accuracy decreased in other baseline methods. This is because the natural motion manifold is not considered in baseline methods, which means the generated adversarial samples are far away from the motion manifold, decreasing the standard accuracy.

To further understand the reason, we plot the deviation distributions of on/off-manifold adversarial samples across three models. The deviation is computed based on the l_2 distance between each clean sample and its corresponding adversarial sample. Since the distributions are similar across the three models, we only shown the distribution on ST-GCN (Fig. 3). From the figure, we can see that there are long tails in both distributions. Further there is clearly more than one mode when combining both distributions. Random Smoothing [42] essentially expands the data distribution homogeneously and symmetrically by a fixed distance at every data point, while TRADES and MART draw perturbations via specific adversary. Their strategy is overly simplified and hence cannot capture the fine-grained structure of the adversarial distribution like MMAT. This shows the necessity of modeling the distributions of on/off manifold samples separately

5.3. Gradient obfuscation evaluation

Gradient obfuscation can potentially lead to failure in defense methods [43], because obfuscated gradients can be circumvented and are not truly robust. Considering that adaptive attack has become the de facto standard for evaluating gradient obfuscation [43,44], we following the adaptive attack criterion [44] to deploy an adaptive attack called EOT-SMART for skeleton-based HAR: in each step, we estimate the expected gradient by averaging the gradients of multiple randomly interpolated samples. Table 8 shows that the robustness under the adaptive attack only slightly worse than original SMART-200, demonstrating that MMAT does not rely on obfuscated gradients.

6. Discussion and conclusion

We proposed the very first black-box adversarial attack method which gives strong performance across datasets, models and attack modes. More broadly, we show, for the first time, the wide existence of on-manifold adversarial samples in skeletal motions. We also proposed a new adversarial training method to achieve simultaneous improvement on accuracy and robustness in general. One limitation is that BASAR relies on an explicit manifold parameterization which is not always feasible in other data types, e.g. videos [45]. This can be mitigated by learning from a large video dataset and use the learned model to do the manifold projection. Finally, BASAR adversarial samples can be theoretically realized by humans because they are on the natural manifold. However, how to attack a system in the real world

using BASAR still depends on the specific setting of the system. In this research, we make the first step to identify the potential danger. In future, we will extend BASAR in other modalities via implicit manifold parameterization.

CRediT authorship contribution statement

Yunfeng Diao: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation. **He Wang:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Funding acquisition, Formal analysis. **Tianjia Shao:** Writing – original draft, Visualization, Formal analysis, Conceptualization. **Yongliang Yang:** Writing – review & editing, Writing – original draft, Funding acquisition, Formal analysis, Conceptualization. **Kun Zhou:** Resources, Funding acquisition, Formal analysis, Conceptualization. **David Hogg:** Writing – review & editing, Validation, Formal analysis. **Meng Wang:** Writing – review & editing, Supervision, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

Appendix A. Supplementary data

Supplementary material related to this article can be found online.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, 2014, arXiv:1312.6199 [cs].

[2] P. Bountakas, A. Zarras, A. Lekidis, C. Xenakis, Defense strategies for adversarial machine learning: A survey, Comp. Sci. Rev. 49 (2023) 100573.

[3] F. Karim, S. Majumdar, H. Darabi, Adversarial attacks on time series, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2021) 3309–3320.

[4] J. Liu, N. Akhtar, A. Mian, Adversarial attack on skeleton-based human action recognition, IEEE Trans. Neural Netw. Learn. Syst. 33 (4) (2022) 1609–1622.

[5] T. Zheng, S. Liu, C. Chen, J. Yuan, B. Li, K. Ren, Towards understanding the adversarial vulnerability of skeleton-based action recognition, 2020, arXiv:2005.07151 [cs].

[6] H. Wang, F. He, Z. Peng, T. Shao, Y. Yang, K. Zhou, D. Hogg, Understanding the robustness of skeleton-based action recognition under adversarial attack, in: CVPR, 2021, pp. 14656–14665.

[7] D. Stutz, M. Hein, B. Schiele, Disentangling adversarial robustness and generalization, in: CVPR, 2019, pp. 6969–6980.

[8] H. Zhang, Y. Yu, J. Jiao, E.P. Xing, L.E. Ghaoui, M.I. Jordan, Theoretically principled trade-off between robustness and accuracy, in: ICML, Vol. 97, 2019, pp. 7472–7482.

[9] Y. Diao, T. Shao, Y.-L. Yang, K. Zhou, H. Wang, BASAR:Black-box attack on skeletal action recognition, in: CVPR, 2021, pp. 7597–7607.

[10] C. Si, Y. Jing, W. Wang, L. Wang, T. Tan, Skeleton-based action recognition with hierarchical spatial reasoning and temporal stack learning network, Pattern Recognit. 107 (2020) 107511.

[11] Y. Li, R. Xia, X. Liu, Learning shape and motion representations for view invariant skeleton-based action recognition, Pattern Recognit. 103 (2020) 107293.

[12] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: AAAI, 2018, pp. 7444–7452.

[13] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, N. Zheng, Semantics-guided neural networks for efficient skeleton-based human action recognition, in: CVPR, 2020, pp. 1109–1118.

[14] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, W. Hu, Channel-wise topology refinement graph convolution for skeleton-based action recognition, in: ICCV, 2021, pp. 13359–13368.

[15] H. Zhou, Q. Liu, Y. Wang, Learning discriminative representations for skeleton based action recognition, in: CVPR, 2023, pp. 10608–10617.

[16] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, F. Lu, Understanding adversarial attacks on deep learning based medical image analysis systems, Pattern Recognit. 110 (2021) 107332.

[17] X. Yang, M. Wang, D. Tao, Person re-identification with metric learning using privileged information, IEEE Trans. Image Process. 27 (2) (2017) 791–805.

[18] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: ICLR, 2015.

[19] A. Kurakin, I. Goodfellow, S. Bengio, Adversarial examples in the physical world, in: ICLR Workshop, 2017.

[20] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: SP, 2017, pp. 39–57.

[21] W. Brendel, J. Rauber, M. Bethge, Decision-based adversarial attacks: Reliable attacks against black-box machine learning models, in: ICLR, 2018.

[22] J. Chen, Q. Gu, RayS: A ray searching method for hard-label adversarial attack, in: KDD, 2020, pp. 1739–1747.

[23] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: ICLR, 2018.

[24] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, Q. Gu, Improving adversarial robustness requires revisiting misclassified examples, in: ICLR, 2020.

[25] Z. Wang, T. Pang, C. Du, M. Lin, W. Liu, S. Yan, Better diffusion models further improve adversarial training, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), ICML, Vol. 202, 2023, pp. 36246–36263.

[26] S. Addepalli, S. Jain, et al., Efficient and effective augmentation strategy for adversarial training, NeurIPS 35 (2022) 1488–1501.

[27] D. Stutz, M. Hein, B. Schiele, Disentangling adversarial robustness and generalization, in: CVPR, 2019, pp. 6976–6987.

[28] Y. Yang, C. Rashtchian, H. Zhang, R.R. Salakhutdinov, K. Chaudhuri, A closer look at accuracy vs. Robustness, in: NeurIPS, 2020.

[29] T. Pang, M. Lin, X. Yang, J. Zhu, S. Yan, Robustness and accuracy could be reconcilable by (proper) definition, in: ICML, 2022, pp. 17258–17277.

[30] S.-Y. Lo, V.M. Patel, Defending against multiple and unforeseen adversarial videos, IEEE Trans. Image Process. (2021).

[31] X. Yang, S. Wang, J. Dong, J. Dong, M. Wang, T.-S. Chua, Video moment retrieval with cross-modal neural architecture search, IEEE Trans. Image Process. 31 (2022) 1204–1216.

[32] J.D. Hedengren, R.A. Shishavan, K.M. Powell, T.F. Edgar, Nonlinear modeling, estimation and predictive control in apmonitor, Comput. Chem. Eng. 70 (2014) 133–148.

[33] Z. Liu, H. Zhang, Z. Chen, Z. Wang, W. Ouyang, Disentangling and unifying graph convolutions for skeleton-based action recognition, in: CVPR, 2020, pp. 140–149.

[34] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, A. Weber, Documentation Mocap Database HDM05, Tech. Rep. CG-2007-2, Universität Bonn, 2007.

[35] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu rgb+ d: A large scale dataset for 3d human activity analysis, in: CVPR, 2016, pp. 1010–1019.

[36] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, Z. Li, UAV-human: A large benchmark for human behavior understanding with unmanned aerial vehicles, CVPR (2021) 16261–16270.

[37] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, A. Zisserman, The kinetics human action video dataset, 2017, arXiv:1705.06950.

[38] P. Gupta, A. Thatipelli, A. Aggarwal, S. Maheshwari, N. Trivedi, S. Das, R.K. Sarvadevabhatla, Quo vadis, skeleton action recognition? Int. J. Comput. Vis. 129 (7) (2021) 2097–2112.

[39] J. Carreira, E. Noland, C. Hillier, A. Zisserman, A short note on the kinetics-700 human action dataset, 2019, arXiv:1907.06987.

[40] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: CVPR, 2015, pp. 1110–1118.

[41] L. Shi, Y. Zhang, J. Cheng, H. Lu, Two-stream adaptive graph convolutional networks for skeleton-based action recognition, in: CVPR, 2019, pp. 12026–12035.

[42] J.M. Cohen, E. Rosenfeld, J.Z. Kolter, Certified adversarial robustness via randomized smoothing, in: ICML, Vol. 97, 2019, pp. 1310–1320.

[43] A. Athalye, N. Carlini, D. Wagner, Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples, in: ICML, PMLR, 2018, pp. 274–283.

[44] F. Tramer, N. Carlini, W. Brendel, A. Madry, On adaptive attacks to adversarial example defenses, NeurIPS 33 (2020) 1633–1645.

[45] X. Yang, F. Feng, W. Ji, M. Wang, T.-S. Chua, Deconfounded video moment retrieval with causal intervention, in: ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1–10.

Yunfeng Diao is an assistant professor at the School of Computer Science and Information Engineering, Hefei University of Technology, China. He received his Ph.D. from Southwest Jiaotong University, China. Previously he was a visiting Ph.D. student at the School of Computing, University of Leeds, UK. His current research interests include computer vision and the security of machine learning.

He Wang is an Associate Professor at the Department of Computer Science, University College London and a Visiting Professor at the University of Leeds. Previously, he was a Turing Fellow, the Director of High-Performance Graphics and Game Engineering at the University of Leeds. He serves as an Associate Editor of Computer Graphics Forum. His research interest is mainly in computer graphics, computer vision and machine learning. He received his Ph.D. from the University of Edinburgh.

Tianjia Shao is a ZJU100 Young Professor in the State Key Laboratory of CAD&CG, Zhejiang University. Previously He was an Assistant Professor (Lecturer in UK) in the School of Computing, University of Leeds, UK. He received his Ph.D. in Computer Science from Institute for Advanced Study, Tsinghua University. His research focuses on the acquisition, creation and processing of 3D contents in Computer Graphics and Computer Vision.

Yong-Liang Yang is a Senior Lecturer at the Department of Computer Science, University of Bath. Previously he was Research Scientist and Post-Doctorate Fellow at King Abdullah University of Science and Technology. He received the B.S. degree and the Ph.D. degree of Computer Science from Tsinghua University. His research interests are broadly in visual computing and interactive techniques.

Kun Zhou received the B.S. degree and Ph.D. degree in computer science from Zhejiang University, in 1997 and 2002, respectively. He is a Cheung Kong professor with the Computer Science Department, Zhejiang University, and the director of the State Key Lab of CAD&CG. Prior to joining Zhejiang University in 2008, he was a leader researcher of the Internet Graphics Group, Microsoft Research Asia. He was named one of the world's top 35 young innovators by MIT Technology Review, in 2011. His research interests are in visual computing, parallel computing, human computer interaction, and virtual reality. He currently serves on the editorial/advisory boards of the IEEE Transactions on Visualization and Computer Graphics and the IEEE Spectrum. He is a fellow of the IEEE.

David Hogg is a Professor of Artificial Intelligence at the University of Leeds. His research is in representation and learning of activities from video. He has been a Chair of the ICT Strategic Advisory Team at the Engineering and Physical Sciences Research Council (EPSRC) in UK. He is Director of the UKRI Centre for Doctoral Training in AI for Medical Diagnosis and Care; and a Co-Director of the Northern Pathology Imaging Co-operative. He has been an associate editor for the IEEE Transactions on Pattern Analysis and Machine Intelligence (2006–2010) and program cochair of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. He is a Fellow of the European Association for Artificial Intelligence (EurAI), a Distinguished Fellow of the British Machine Vision Association, and a Fellow of the International Association for Pattern Recognition.

Meng Wang is a professor at Hefei University of Technology, China. His current research interests include multimedia content analysis, computer vision, and pattern recognition. He received paper prizes or awards from ACM MM 2009 (Best Paper Award), ACM MM 2010 (Best Paper Award), MMM 2010 (Best Paper Award), SIGIR 2015 (Best Paper Honorable Mention), IEEE TMM 2015 and 2016 (Prize Paper Award Honorable Mention) and ACM TOMM 2018 (Nicolas D. Georganas Best Paper Award), etc. He currently serves on the editorial/advisory boards of IEEE TPAMI, IEEE TMM, IEEE TNNLS, etc. He is a fellow of IEEE and IAPR.