UNIVERSITY OF LEEDS

# The Universal Vulnerability of Human Action Recognition Classifiers and Potential Solutions

**when adversarial robustness meets computer graphics**

**He Wang**

UNIVERSITY OF LEEDS

# Introduction

## Deep neural networks are extremely popular

- A wide range of applications, e.g. object recognition, activity recognition, etc.
- Extremely vulnerable to malicious perturbations a.k.a adversarial attack
  - Attacks on training data, testing data, training process, etc.
- Adversarial attack has emerged as a new field
  - Euclidean data, graphs, etc.
  - On image, video, finance, etc.
  - On security/safety related tasks, e.g. self-driving cars

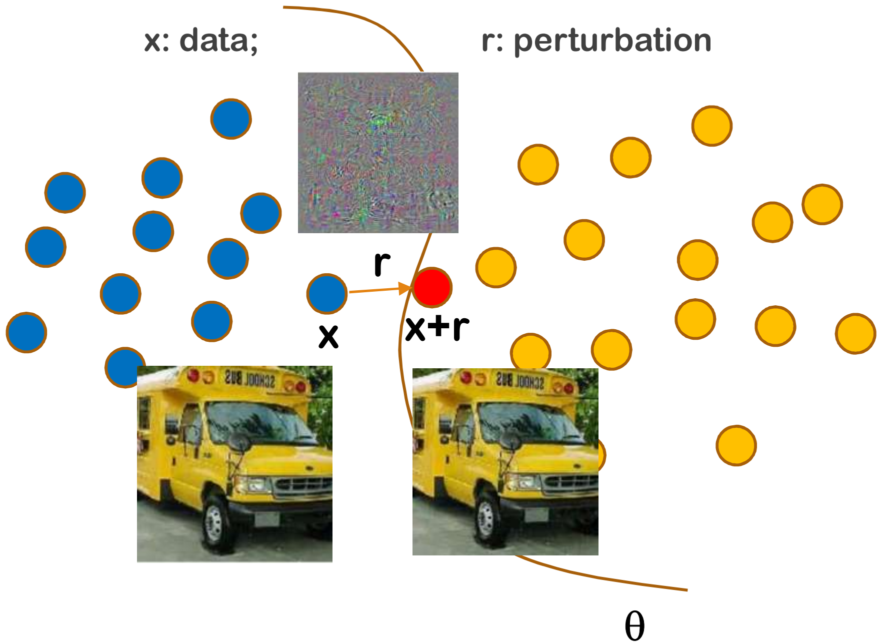
## We examine skeleton-based activity recognition

- A key data type in many applications
- Unexplored in adversarial attack and defense
- A newly found niche with our contributions: new attack and defense methods

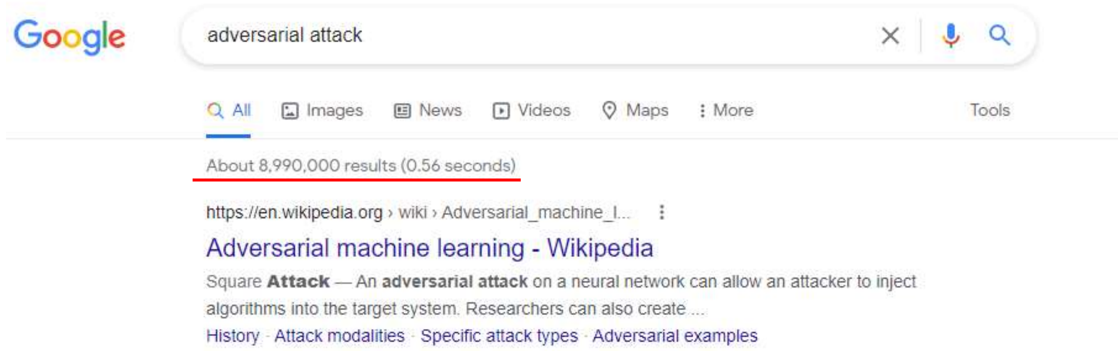
# Background

## Adversarial Attack

- Strategically computed perturbations




A binary classifier $g_\theta$ , $\theta$: network weights




1. Destructive to machine intelligence->fool AI
2. Imperceptible to humans->fool humans

UNIVERSITY OF LEEDS

# Existing work

| Method | Training Process | Testing Process | Victim Model | Surrogate |
|---|---|---|---|---|
| White-box attack | Sometimes | Yes | Yes | No |
| Black-box attack | No | Yes | No | Sometimes |
| White-box defense | Yes | Sometimes | Yes | No |
| Black-box defense | Sometimes | Yes | No | Sometimes |

# Human Activity Recognition

Human Activity Recognition (HAR):An important type of time-series data

- Data types: image, video, skeleton, etc.
- We focus on 3D skeleton based classifiers
  - Robust to lighting, occlusion, ambiguity, etc.


Challenges:

- Not well studied in the context of adversarial attack and defense
- Low-dimensionality
  - less than 100 Dofs per frame which restricts the attack
- Perceptual Sensitivity
  - any perturbation on single joints or single frames is trivially identifiable.

**He Wang, Feixiang He, Zhexi Peng, Tianjia Shao, Yongliang Yang, Kun Zhou and David Hogg, Understanding the Robustness of Skeleton-based Action Recognition under Adversarial Attack, CVPR 2021**

UNIVERSITY OF LEEDS

# SMART

## SMART: Skeletal Motion Action Recognition aTtack , a while-box attack

$$\arg\min_{\hat{q}} w\boxed{L_c(q, \hat{q})} + (1-w)\boxed{L_p(q, \hat{q})}$$

classification loss perceptual loss

dynamics loss

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

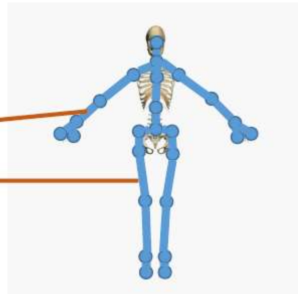
$$l_{dyn} = \sum_{n=0}^{\infty} \beta_n ||(q^n - \hat{q}^n)||_2^2 \text{ where} \sum_{n=0}^{\infty} \beta_n = 1$$

Derivative matching to keep the motion dynamics

Bone-length loss

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

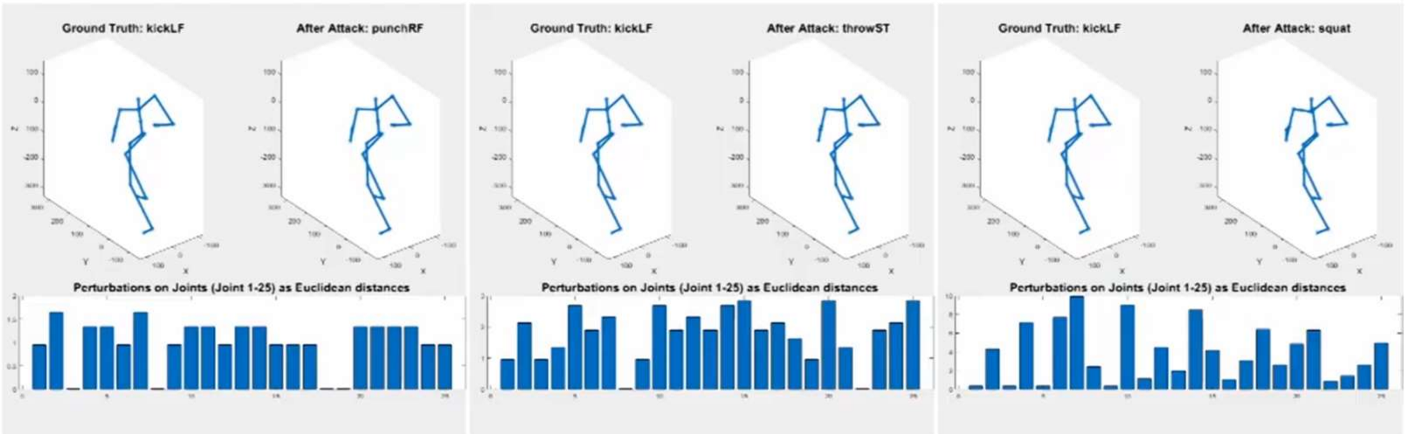
$$l_{bl} = ||Bl(q) - Bl(\hat{q})||_2^2 = \frac{1}{M} \sum_{i=1}^{M} ||Bl(q_i) - Bl(\hat{q}_i)||_2^2$$



**He Wang, Feixiang He, Zhexi Peng, Tianjia Shao, Yongliang Yang, Kun Zhou and David Hogg, Understanding the Robustness of Skeleton-based Action Recognition under Adversarial Attack, CVPR 2021**

# Experiments

Target model: **HRNN**; Dataset: **HDM05**




Attacking Strategy: **AB** Attacking Strategy: **ABN** Attacking Strategy: **SA**

**He Wang, Feixiang He, Zhexi Peng, Tianjia Shao, Yongliang Yang, Kun Zhou and David Hogg, Understanding the Robustness of Skeleton-based Action Recognition under Adversarial Attack, CVPR 2021**

# Experiments

| Model/Data | Anything-but (AB) Attack HDM05 | MHAD | NTU | Anything-but-N Attack HDM05 | MHAD | NTU | Specified Attack (SA) 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 |

Transfer-based black-box
- Attack a surrogate model via SMART, then use the results to attack other models

Strict Perceptual Study
- SMART can easily fool human eyes

He Wang, Feixiang He, Zhexi Peng, Tianjia Shao, Yongliang Yang, Kun Zhou and David Hogg, Understanding the Robustness of Skeleton-based Action Recognition under Adversarial Attack, CVPR 2021

UNIVERSITY OF LEEDS

# Experiments

Ground Truth: kickLS

After Attack: throwFR | After Attack: grabFR | After Attack: jogOP

**CIASA** | **IAA** | **SMART**

He Wang, Feixiang He, Zhexi Peng, Tianjia Shao, Yongliang Yang, Kun Zhou and David Hogg, Understanding the Robustness of Skeleton-based Action Recognition under Adversarial Attack, CVPR 2021

UNIVERSITY OF LEEDS

# Summary

The first while-box attack on skeleton-based action recognition

- High success rate
- Can do transfer-based black-box attack
- Highly unperceivable to humans
- Existing human activity classifiers are very vulnerable

He Wang, Feixiang He, Zhexi Peng, Tianjia Shao, Yongliang Yang, Kun Zhou and David Hogg, Understanding the Robustness of Skeleton-based Action Recognition under Adversarial Attack, CVPR 2021

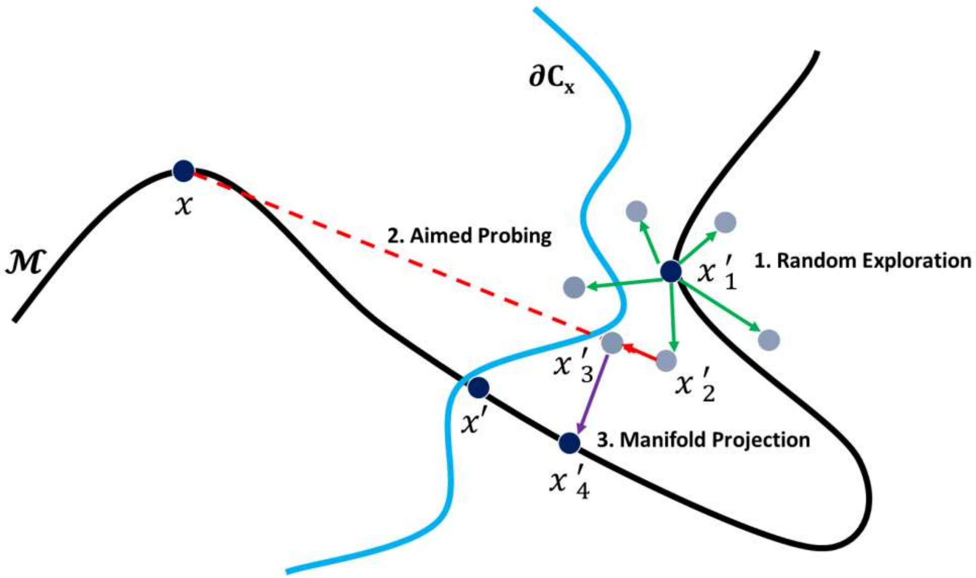
# BASAR (Black-box Attack on Skeletal Action Recognition)

SMART is a white-box approach
- Needs to know the full details of the trained network
- What if the network details are unavailable?

The first black-box attack method on HAR
- Also needs to address low-dimensionality and perceptual sensitivity
- Existence of on-manifold adversarial examples
  - Existing research believes adversarial examples are off data manifold
  - We show the wide existence of on-manifold adversarial samples

**Yunfeng Diao, Tianjia Shao, Yongliang Yang, Kun Zhou and He Wang, BASAR:Black-box Attack on Skeletal Action Recognition, CVPR 2021**

# BASAR

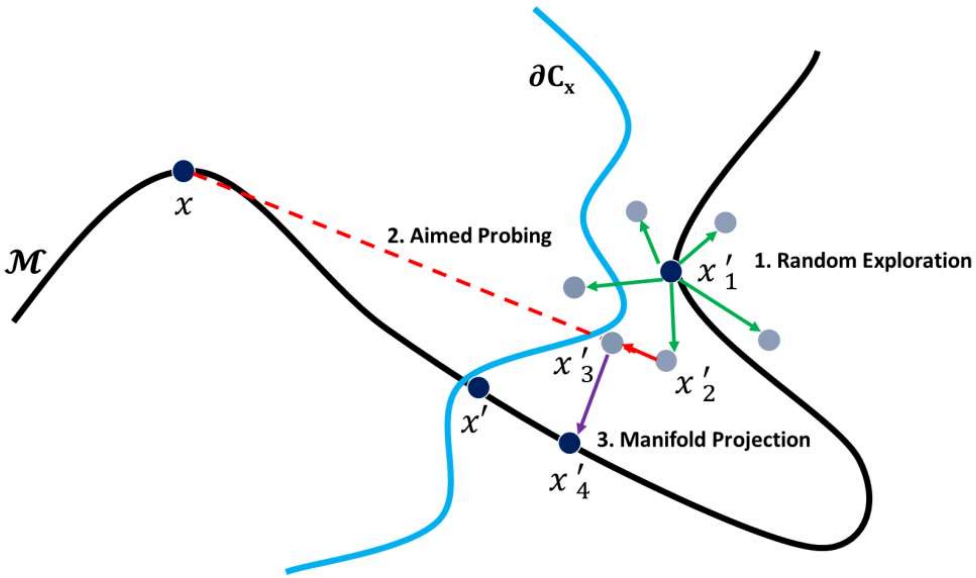



$$
\begin{aligned}
&\text{minimize} \quad L(\mathbf{x}, \mathbf{x}') \\
&\text{subject to} \quad \mathbf{x}' \in [0,1]^{m \times n}, \mathbf{x}' \in \mathcal{M} \\
&C_{\mathbf{x}'} = c \text{ (targeted) or } C_{\mathbf{x}'} \neq C_{\mathbf{x}} \text{ (untargeted).}
\end{aligned}
$$

**Yunfeng Diao, Tianjia Shao, Yongliang Yang, Kun Zhou and He Wang, BASAR:Black-box Attack on Skeletal Action Recognition, CVPR 2021**

UNIVERSITY OF LEEDS

# BASAR




## Random Exploration

$$\widetilde{\mathbf{x}} = \mathbf{x}' + \mathbf{W}\Delta,$$

$$\text{where } \Delta_* = \mathbf{R}_* - (\mathbf{R}_*^T\mathbf{d}_*)\mathbf{d}_*,\ \mathbf{d}_* = \frac{\mathbf{x}_* - \mathbf{x}'_*}{\|\mathbf{x}_* - \mathbf{x}'_*\|},$$

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

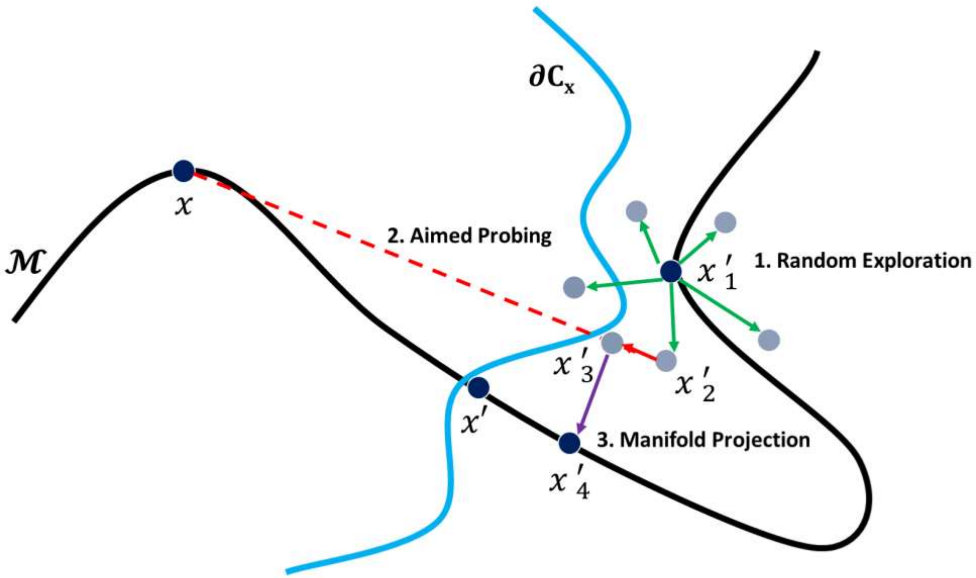
## Aimed Probing

$$\widetilde{\mathbf{x}} = \mathbf{x}' + \beta(\mathbf{x} - \mathbf{x}')$$

Yunfeng Diao, Tianjia Shao, Yongliang Yang, Kun Zhou and He Wang, BASAR:Black-box Attack on Skeletal Action Recognition, CVPR 2021

UNIVERSITY OF LEEDS

# BASAR




**Manifold Projection**

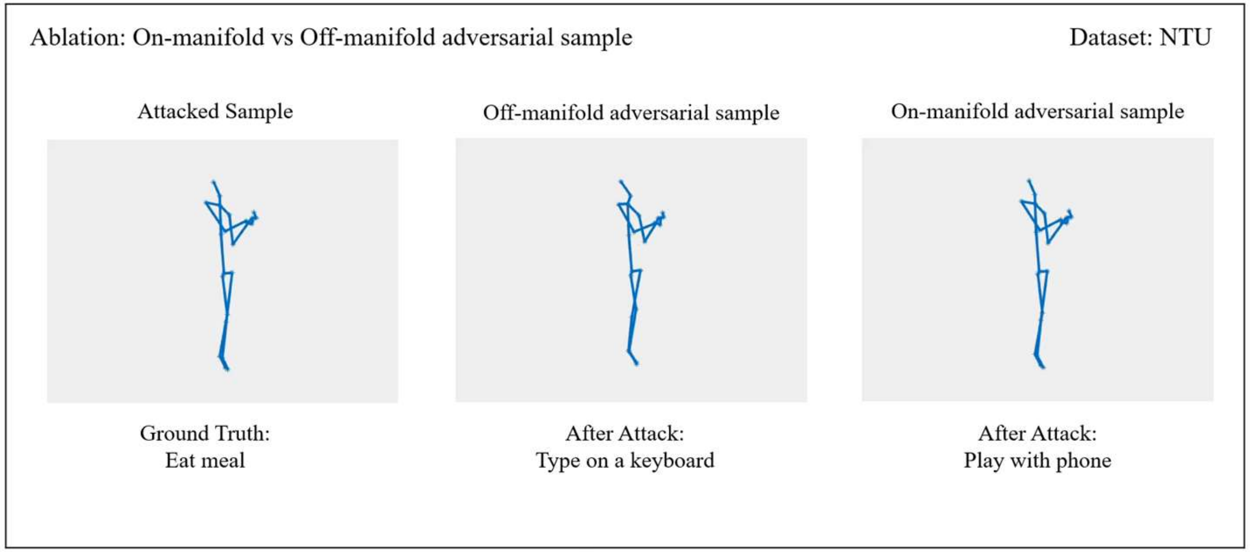
$$\min_{\mathbf{x}'} \; L(\widetilde{\mathbf{x}}, \mathbf{x}') + wL(\ddot{\mathbf{x}}, \ddot{\mathbf{x}}')$$

subject to $B'_i = B_i$ and $\theta_i^{\min} \leq \theta'_i \leq \theta_i^{\max}$

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

**Yunfeng Diao, Tianjia Shao, Yongliang Yang, Kun Zhou and He Wang, BASAR:Black-box Attack on Skeletal Action Recognition, CVPR 2021**

# Experiments




Yunfeng Diao, Tianjia Shao, Yongliang Yang, Kun Zhou and He Wang, BASAR:Black-box Attack on Skeletal Action Recognition, CVPR 2021

UNIVERSITY OF LEEDS

# Summary

## The first black-box attack on skeleton-based action recognition

- High success rate
- Show wide existence of on-manifold adversarial samples, for the first time
- Highly unperceivable attacks

## On-going research

- No-box attack: no victim details, no query to the victim, no training data or labels, no surrogate classifiers
  - White-box: needs to know the details of the victim
  - Black-box: needs to access the training data and labels and query the victim

**Yunfeng Diao, Tianjia Shao, Yongliang Yang, Kun Zhou and He Wang, BASAR:Black-box Attack on Skeletal Action Recognition, CVPR 2021**

UNIVERSITY OF LEEDS

# Defense

## Defending Black-box Skeleton-based Human Activity Classifiers

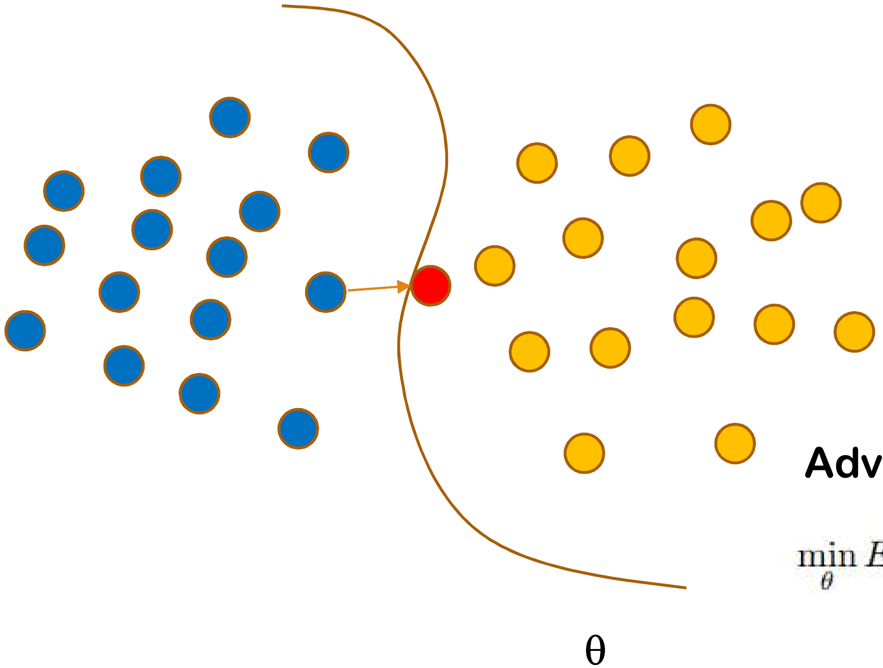
## No defense method has been designed for HAR

- Existing defense methods are ineffective
  - They are mostly designed for static data, e.g. images
  - They do not consider dynamics in time-series
  - They have intrinsic trade-offs between accuracy and robustness
- Typical methods
  - Adversarial Training (AT)
  - Randomized Smoothing (RS)

## Our method: Bayesian Energy-based Adversarial Training (BEAT)

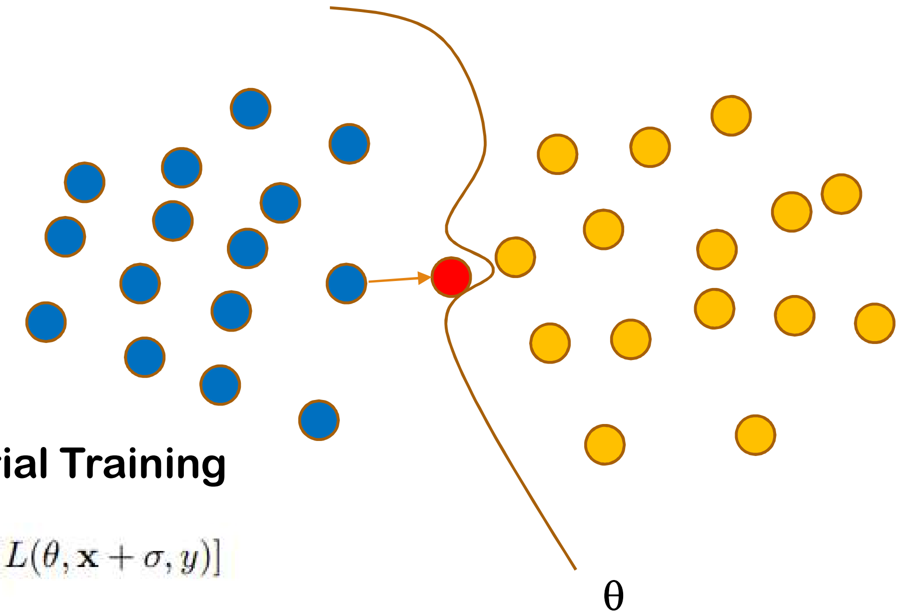
He Wang, Yunfeng Diao, Zichang Tan and Guodong Guo, Defending Black-box Skeleton-based Human Activity Classifiers, AAAI 2023

# Motivation



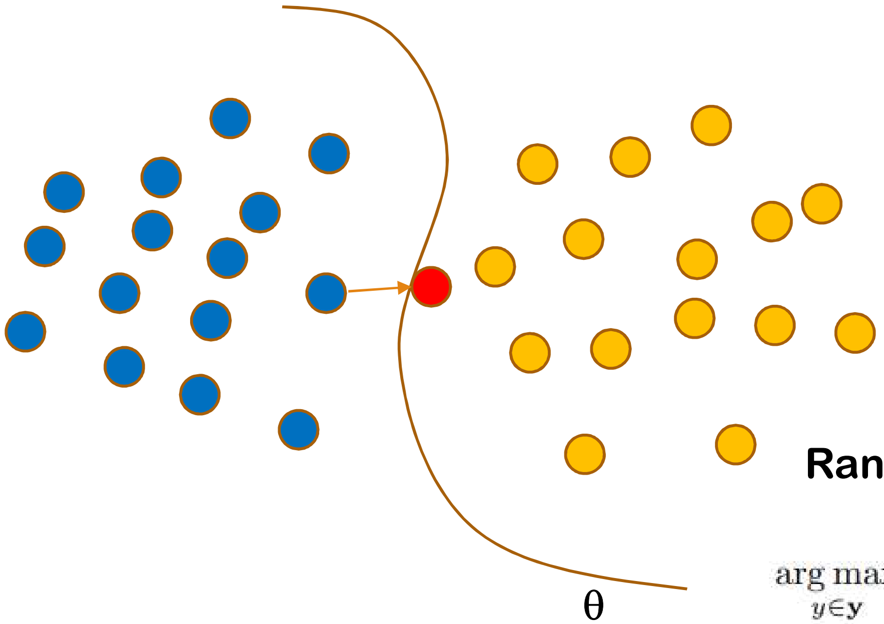

**Adversarial Training**

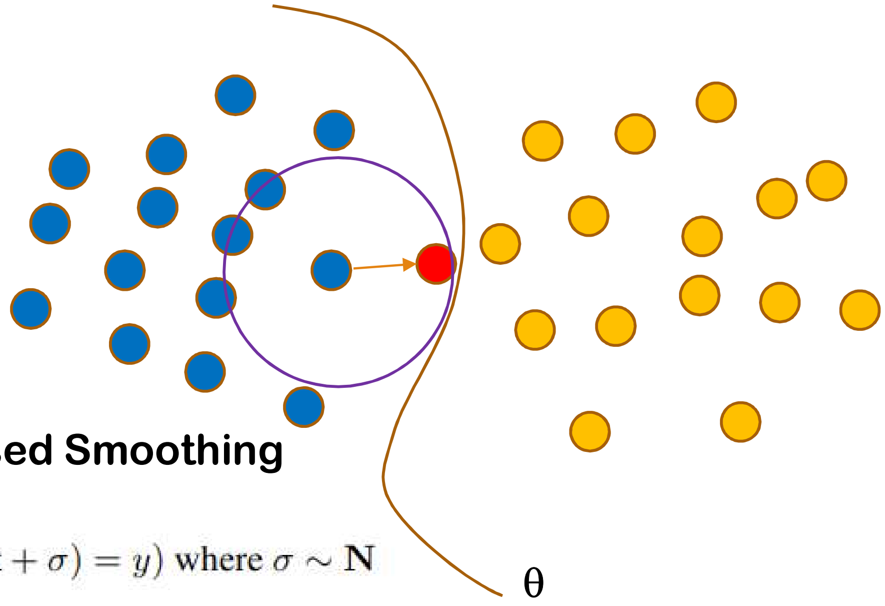
$$\min_{\theta} E_{\mathbf{x}}[\max_{\sigma \in S} L(\theta, \mathbf{x} + \sigma, y)]$$




**He Wang, Yunfeng Diao, Zichang Tan and Guodong Guo, Defending Black-box Skeleton-based Human Activity Classifiers, AAAI 2023**

# Motivation



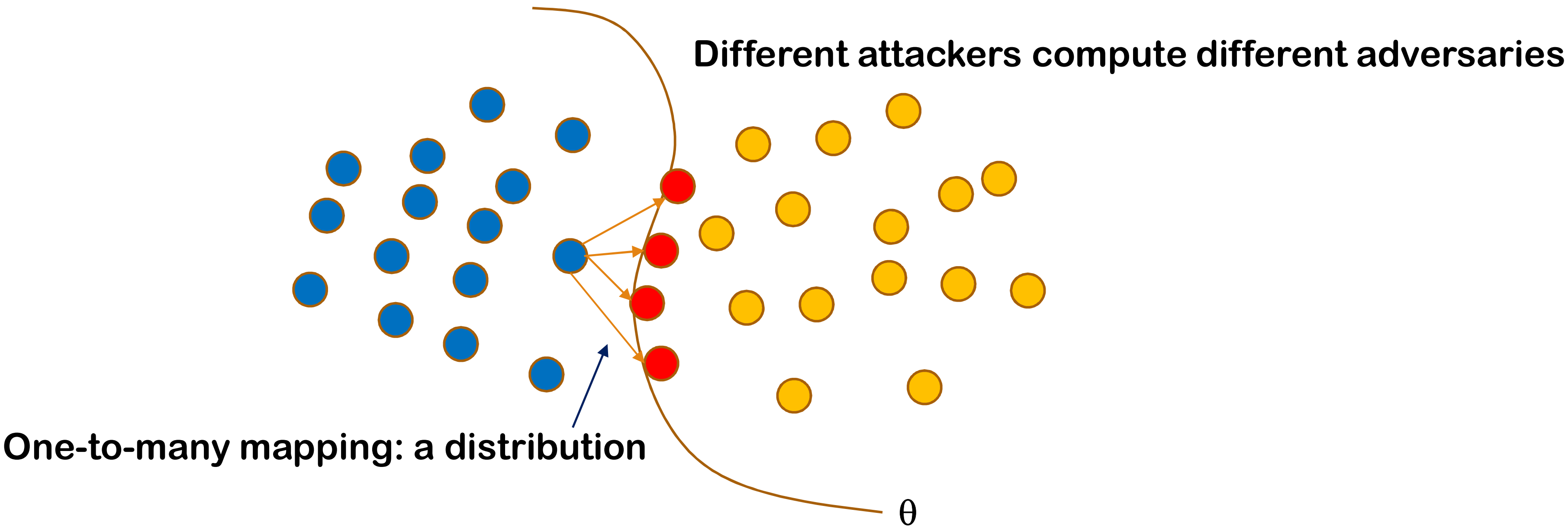

**Randomised Smoothing**

$$\arg\max_{y \in \mathbf{y}} p(g_\theta(\mathbf{x}+\sigma) = y) \text{ where } \sigma \sim \mathbf{N}$$

**He Wang, Yunfeng Diao, Zichang Tan and Guodong Guo, Defending Black-box Skeleton-based Human Activity Classifiers, AAAI 2023**

# Motivation

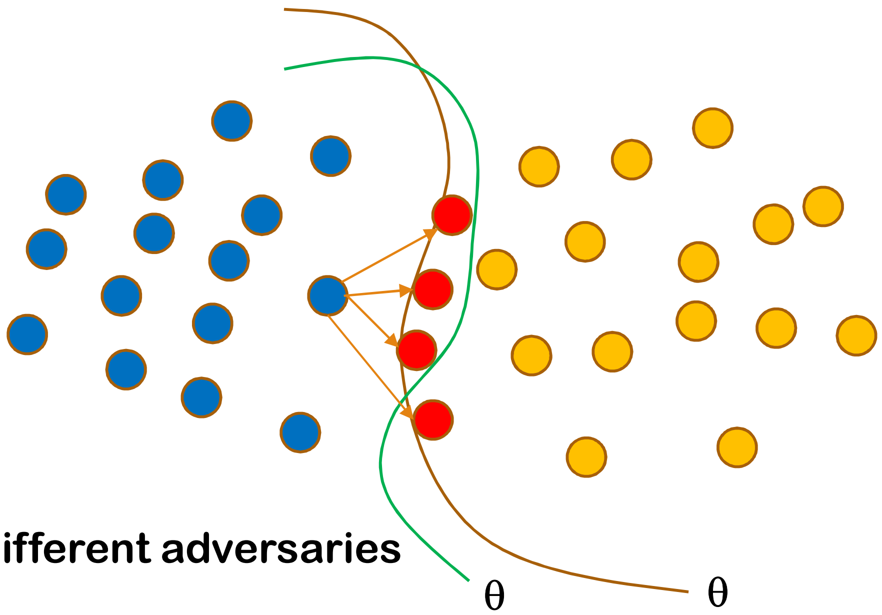
- The necessity to capture the whole adversarial distribution




He Wang, Yunfeng Diao, Zichang Tan and Guodong Guo, Defending Black-box Skeleton-based Human Activity Classifiers, AAAI 2023

# Motivation

- The necessity to capture the distribution of all classifiers




**Different classifiers resist different adversaries**

He Wang, Yunfeng Diao, Zichang Tan and Guodong Guo, Defending Black-box Skeleton-based Human Activity Classifiers, AAAI 2023

# BEAT

UNIVERSITY OF LEEDS

Jointly model data, adversarial and classifier

data adversaries

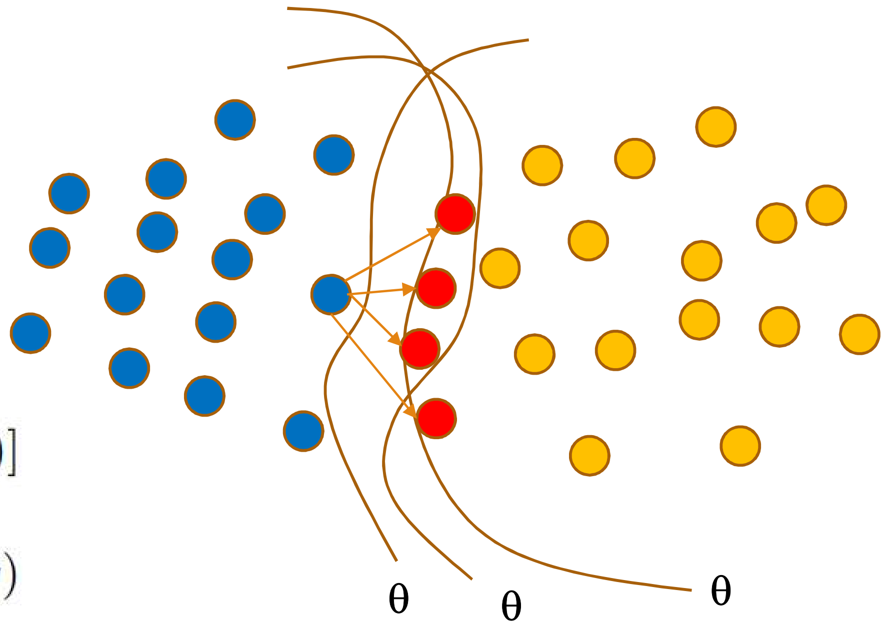
Network weight $p(\theta, \mathbf{x}, \tilde{\mathbf{x}}, y)$ label

Prediction by Bayesian Model Averaging:

Test data

$$p(y'|\mathbf{x}', \mathbf{x}, \tilde{\mathbf{x}}, y) = E_{\theta \sim p(\theta)}[p(y'|\mathbf{x}', \mathbf{x}, \tilde{\mathbf{x}}, y, \theta)]$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} p(y'|\mathbf{x}', \theta_i), \theta \sim p(\theta|\mathbf{x}, \tilde{\mathbf{x}}, y)$$




He Wang, Yunfeng Diao, Zichang Tan and Guodong Guo, Defending Black-box Skeleton-based Human Activity Classifiers, AAAI 2023

# BEAT

UNIVERSITY OF LEEDS

## Jointly model data, adversarial distribution and classifier

data

adversaries

$$p(\theta, \mathbf{x}, \tilde{\mathbf{x}}, y)$$

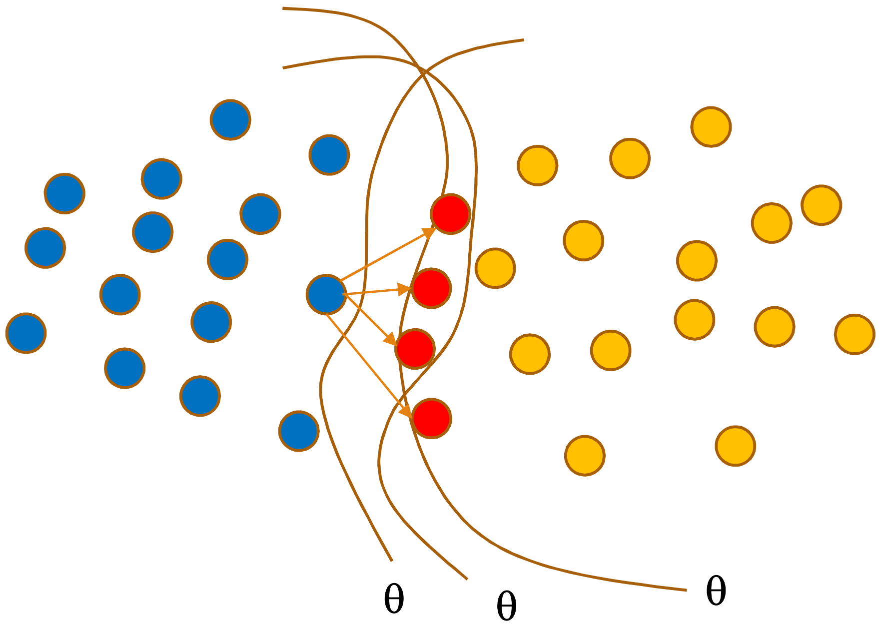
Network weight

label

$$p(\theta, \mathbf{x}, \tilde{\mathbf{x}}, y) = p(\mathbf{x}, \tilde{\mathbf{x}}, y|\theta)p(\theta)$$

$$p_\theta(\mathbf{x}, \tilde{\mathbf{x}}, y) \quad = \quad p_\theta(\tilde{\mathbf{x}}|\mathbf{x}, y)p_\theta(\mathbf{x}, y)$$

Clean-adversarial joint distribution




He Wang, Yunfeng Diao, Zichang Tan and Guodong Guo, Defending Black-box Skeleton-based Human Activity Classifiers, AAAI 2023

UNIVERSITY OF LEEDS

# BEAT

## Novelty 1: Bayesian treatment on the adversaries->a clean-adversarial distribution

$$p_\theta(\mathbf{x}, y) = \frac{exp(g_\theta(\mathbf{x})[y])}{Z(\theta)}$$

**Some distance function**

$$p_\theta(\mathbf{x}, \tilde{\mathbf{x}}, y) = \frac{exp\{g_\theta(\mathbf{x})[y] + g_\theta(\tilde{\mathbf{x}})[y] - \lambda d(\mathbf{x}, \tilde{\mathbf{x}})\}}{Z(\theta)}$$

**Correctly classifying clean data**

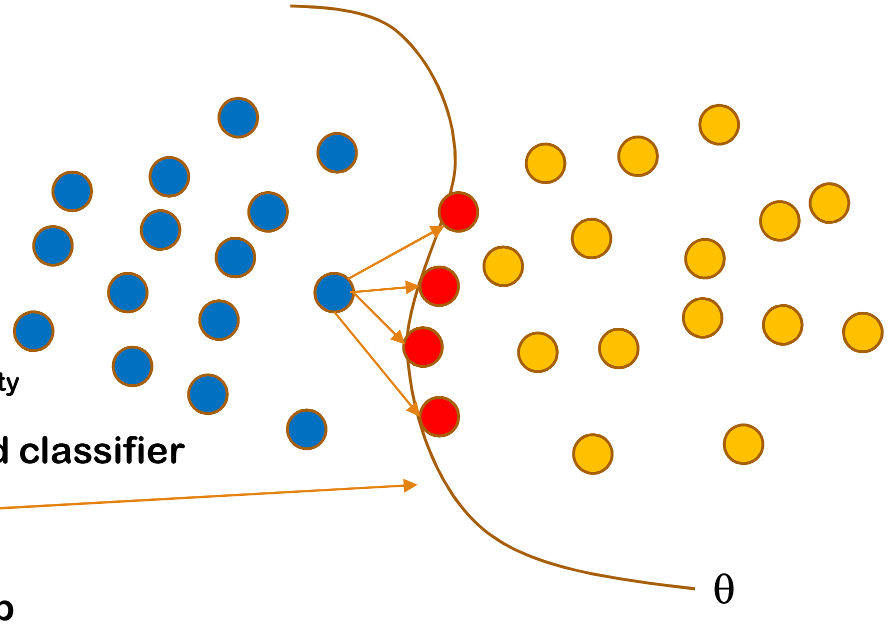
**Correctly classifying adversaries**

**High clean-adversary similarity**

$$p_\theta(\mathbf{x}, \tilde{\mathbf{x}}, y) = p_\theta(\tilde{\mathbf{x}}|\mathbf{x}, y)p_\theta(\mathbf{x}, y)$$

**a normal energy-based classifier**

$$p_\theta(\tilde{\mathbf{x}}|\mathbf{x}, y) = \frac{p_\theta(\mathbf{x}, \tilde{\mathbf{x}}, y)}{p_\theta(\mathbf{x}, y)} = exp\{g_\theta(\tilde{\mathbf{x}})[y] - \lambda d(\mathbf{x}, \tilde{\mathbf{x}})\}$$




**Capturing the one-to-many clean-adversarial relationship**

**He Wang, Yunfeng Diao, Zichang Tan and Guodong Guo, Defending Black-box Skeleton-based Human Activity Classifiers, AAAI 2023**

# BEAT

## Novelty 1: Bayesian treatment on the adversaries->a clean-adversarial distribution

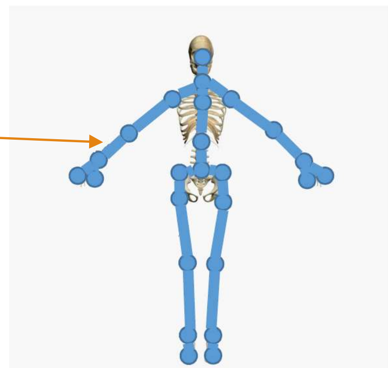
$$p_\theta(\mathbf{x},\tilde{\mathbf{x}},y) = \frac{exp\{g_\theta(\mathbf{x})[y] + g_\theta(\tilde{\mathbf{x}})[y] - \lambda d(\mathbf{x},\tilde{\mathbf{x}})\}}{Z(\theta)}$$

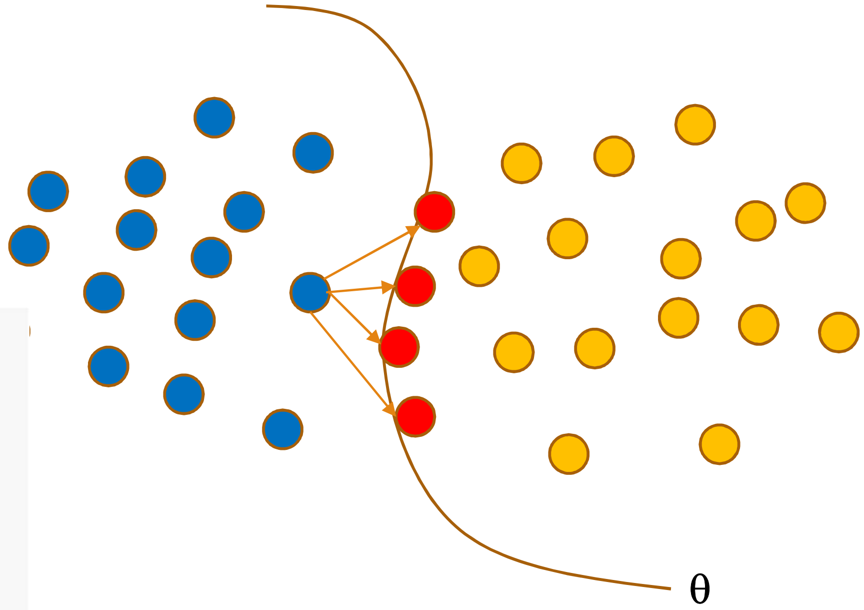
**Measure of similarity between adversaries and clean data**
**Should be domain-specific**

Bone length constraints

$$d(\mathbf{x},\tilde{\mathbf{x}}) = \frac{1}{MB}\sum ||BL(\mathbf{x}) - BL(\tilde{\mathbf{x}})||_p^2 + \frac{1}{MJ}\sum ||q_{m,j}^k(\mathbf{x}) - \tilde{q}_{m,j}^k(\tilde{\mathbf{x}})||_p^2$$

Joint position, velocity, acceleration, etc.






**He Wang, Yunfeng Diao, Zichang Tan and Guodong Guo, Defending Black-box Skeleton-based Human Activity Classifiers, AAAI 2023**

# BEAT

UNIVERSITY OF LEEDS

## Novelty 2: A Bayesian treatment on the classifier

$$p_\theta(\mathbf{x}, \tilde{\mathbf{x}}, y) = \frac{exp\{g_\theta(\mathbf{x})[y] + g_\theta(\tilde{\mathbf{x}})[y] - \lambda d(\mathbf{x}, \tilde{\mathbf{x}})\}}{Z(\theta)}$$

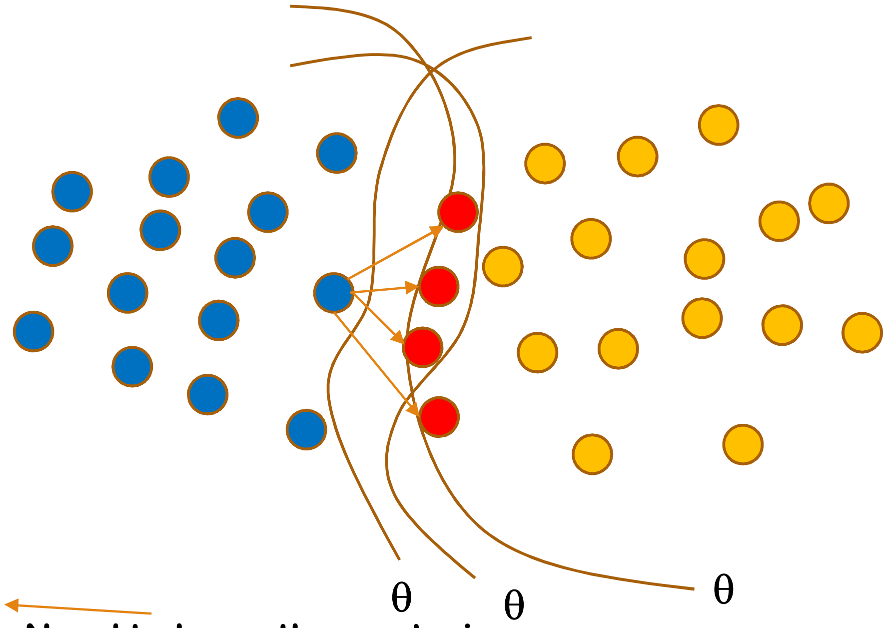
$$p(\theta, \mathbf{x}, \tilde{\mathbf{x}}, y) = p(\mathbf{x}, \tilde{\mathbf{x}}, y|\theta)p(\theta)$$

**Prediction by Bayesian Model Averaging:**

$$p(y'|\mathbf{x}', \mathbf{x}, \tilde{\mathbf{x}}, y) = E_{\theta\sim p(\theta)}[p(y'|\mathbf{x}', \mathbf{x}, \tilde{\mathbf{x}}, y, \theta)]$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} p(y'|\mathbf{x}', \theta_i), \theta \sim p(\theta|\mathbf{x}, \tilde{\mathbf{x}}, y)$$

**Need to learn the posterior**




**He Wang, Yunfeng Diao, Zichang Tan and Guodong Guo, Defending Black-box Skeleton-based Human Activity Classifiers, AAAI 2023**

# BEAT

Novelty 2: A Bayesian treatment on the classifier

However, it is hard to learn $p(\theta, \mathbf{x}, \tilde{\mathbf{x}}, y) = p(\mathbf{x}, \tilde{\mathbf{x}}, y|\theta)p(\theta)$

- The posterior space is too big, sampling is slow
  - millions of parameters in $p(\theta|\mathbf{x}, \tilde{\mathbf{x}}, y)$
- The classifier needs to be retained, inconvenient or impossible
  - Big models are pre-trained and shared
  - Do not have machines or too slow to re-train the model
  - Do not have access to the training data
- Retraining might undermine feature representation, e.g. for other tasks in multi-task learning

He Wang, Yunfeng Diao, Zichang Tan and Guodong Guo, Defending Black-box Skeleton-based Human Activity Classifiers, AAAI 2023

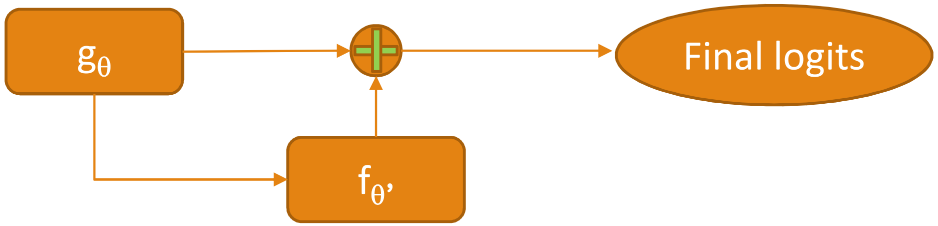
# BEAT

## Novelty 3: A post-train strategy, append a small network f (2-layer MLP)

**Original logits**

$$\text{logits} = f_{\theta'}(\phi(\mathbf{x})) + g_\theta(\mathbf{x})$$

**$\theta$ is pre-trained, we only learn $\theta'$**




1. No re-training needed
2. f is much smaller than g, so training is fast
3. The classifier can be a black-box

He Wang, Yunfeng Diao, Zichang Tan and Guodong Guo, Defending Black-box Skeleton-based Human Activity Classifiers, AAAI 2023

# Experiments

Dataset: HDM05, NTU60 and NTU120

Victim classifier: ST-GCN, CTR-GCN, SGN, MS-G3D

Attackers: SMART, CIASA, BASAR, EOT

Baseline defense methods: Randomized Smoothing, Smart-AT, TRADES, MART

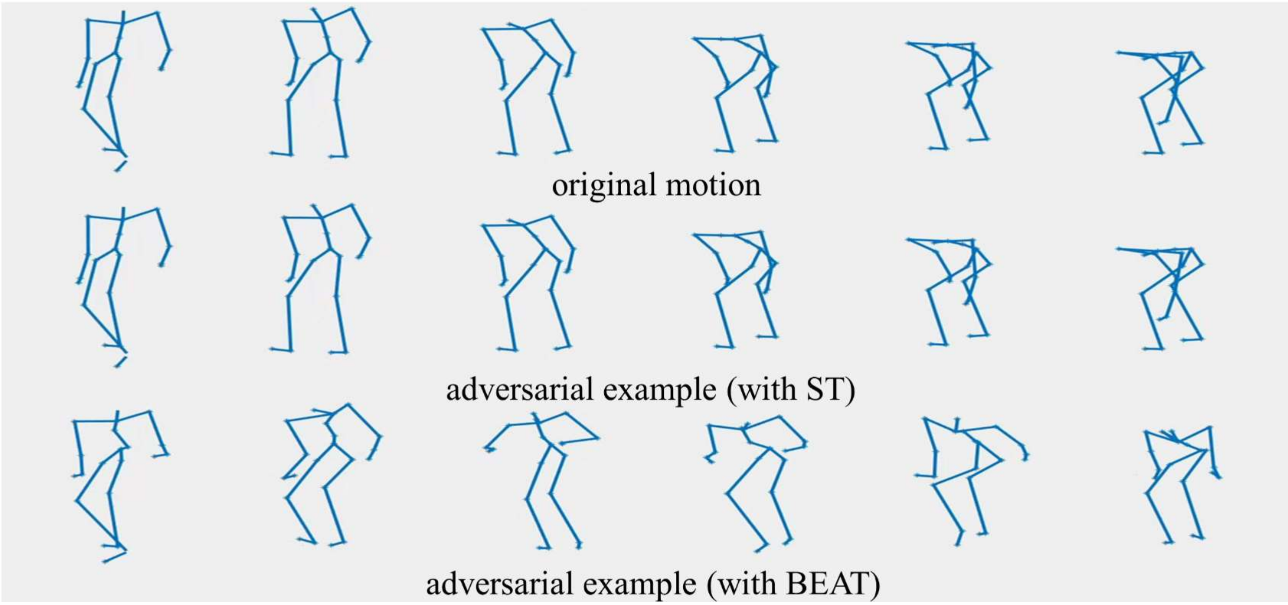
Experimental setting:

- Under white-box attack (SMART and CIASA)
- Under black-box attack (BASAR)
- Under stochastic attack (EOT)

He Wang, Yunfeng Diao, Zichang Tan and Guodong Guo, Defending Black-box Skeleton-based Human Activity Classifiers, AAAI 2023

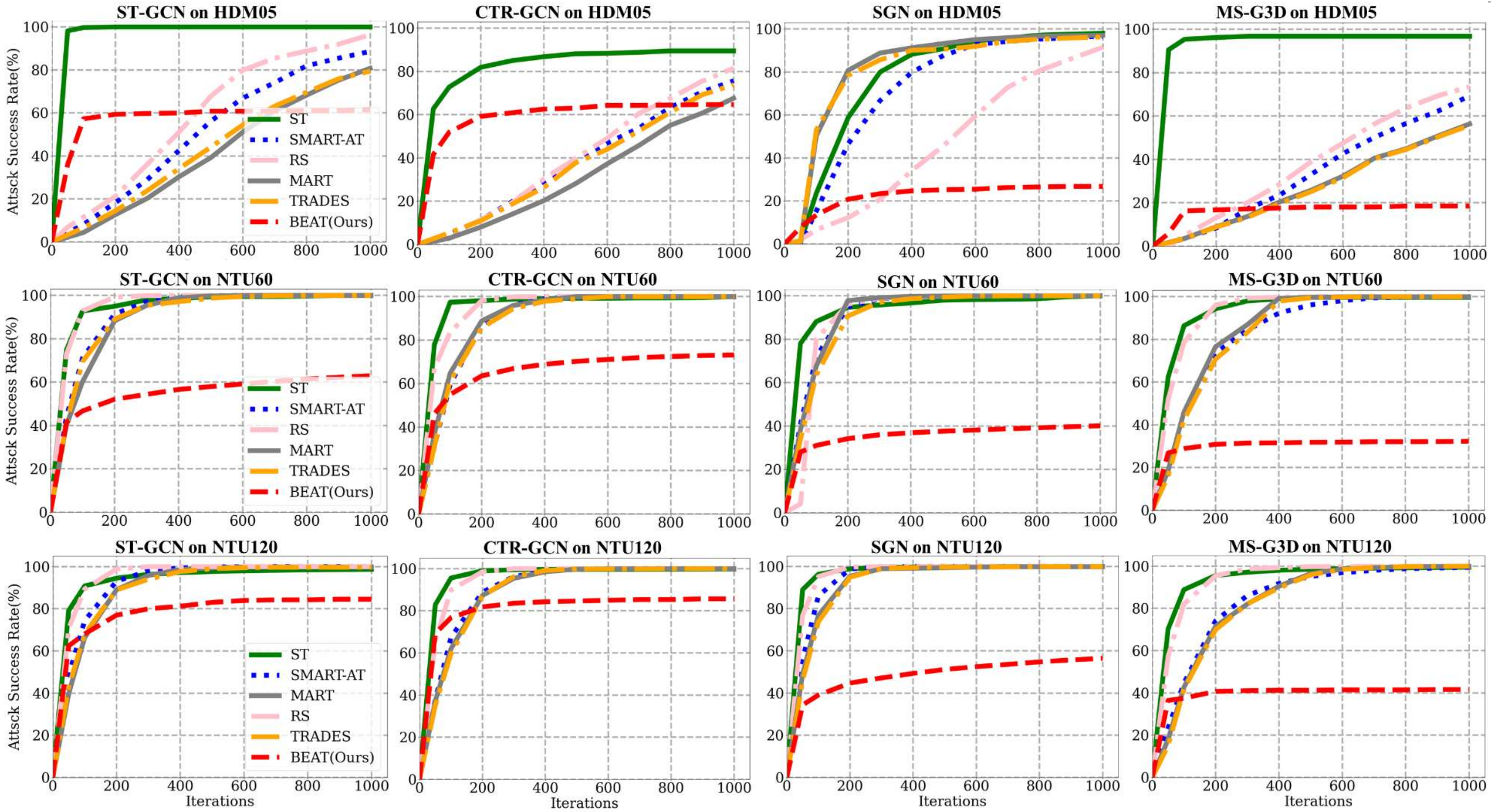
# Experiments

**Large perturbation has to be used to fool a classifier trained on BEAT**

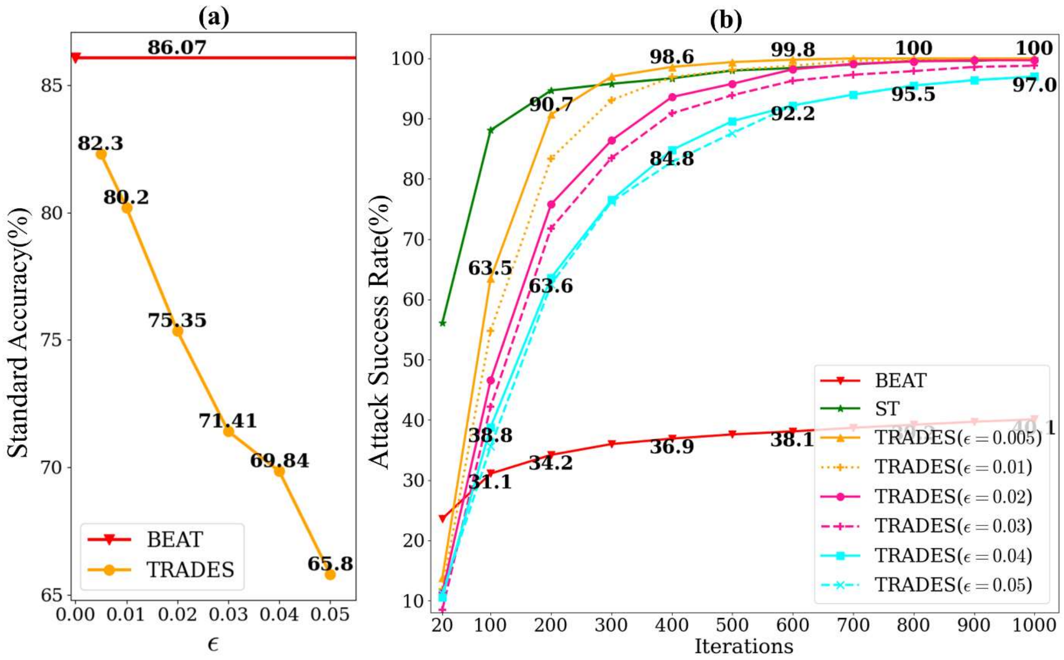



**He Wang, Yunfeng Diao, Zichang Tan and Guodong Guo, Defending Black-box Skeleton-based Human Activity Classifiers, AAAI 2023**

# Experiments



# Experiments




- **Comparisons with TRADES with different perturbation budget epsilon on NTU60 with SGN. (a): standard accuracy vs. epsilon; (b): results against SMART with 20 to 1000 iterations.**

# Summary

The first defense method on skeleton-based action recognition

- High robustness increase
- Fast train, black-box nature
- Effective against various attackers, datasets and classifiers

He Wang, Yunfeng Diao, Zichang Tan and Guodong Guo, Defending Black-box Skeleton-based Human Activity Classifiers, AAAI 2023

# Our broad effort

- The first white-box attack, SMART
  - He Wang, Feixiang He, Zhexi Peng, Tianjia Shao, Yongliang Yang, Kun Zhou and David Hogg, Understanding the Robustness of Skeleton-based Action Recognition under Adversarial Attack, CVPR 2021
- The first black-box attack, BASAR
  - Yunfeng Diao, Tianjia Shao, Yongliang Yang, Kun Zhou and He Wang, BASAR:Black-box Attack on Skeletal Action Recognition, CVPR 2021
- The first black-box defense, BEAT
  - He Wang, Yunfeng Diao, Zichang Tan and Guodong Guo, Defending Black-box Skeleton-based Human Activity Classifiers, AAAI 2023
- A new black-box defense
  - Yunfeng Diao, He Wang, Tianjia Shao, Yong-Liang Yang, Kun Zhou, David Hogg, Understanding the Vulnerability of Skeleton-based Human Activity Recognition via Black-box Attack, arxiv 2022 (under review).
- Resources
  - http://drhewang.com/publications.html

UNIVERSITY OF LEEDS

# Thanks to collaborators and funders

UNIVERSITY OF LEEDS


He Wang


Feixiang He


David C Hogg


Zhexi Peng


Tianjia Shao


Kun Zhou

SOUTHWEST JIAOTONG UNIVERSITY

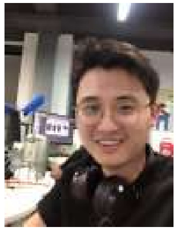
HEFEI UNIVERSITY OF TECHNOLOGY

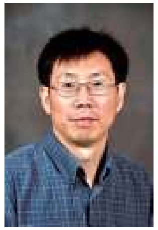
UNIVERSITY OF BATH


Yunfeng Diao


Yongliang Yang

Baidu Research


Zichang Tan


Guodong Guo

CROWDDNA

http://crowddna.eu/

European Commission

Horizon 2020
European Union funding
for Research & Innovation

# Thanks for listening

He Wang