Adversarial Attack on Skeleton-Based Human Action Recognition

Jian Liu, *Member, IEEE*, Naveed Akhtar, *Member, IEEE*, and Ajmal Mian, *Senior Member, IEEE*

***Abstract*— Deep learning models achieve impressive performance for skeleton-based human action recognition. Graph convolutional networks (GCNs) are particularly suitable for this task due to the graph-structured nature of skeleton data. However, the robustness of these models to adversarial attacks remains largely unexplored due to their complex spatiotemporal nature that must represent sparse and discrete skeleton joints. This work presents the first adversarial attack on skeleton-based action recognition with GCNs. The proposed targeted attack, termed constrained iterative attack for skeleton actions (CIASA), perturbs joint locations in an action sequence such that the resulting adversarial sequence preserves the temporal coherence, spatial integrity, and the anthropomorphic plausibility of the skeletons. CIASA achieves this feat by satisfying multiple physical constraints and employing spatial skeleton realignments for the perturbed skeletons along with regularization of the adversarial skeletons with generative networks. We also explore the possibility of semantically imperceptible localized attacks with CIASA and succeed in fooling the state-of-the-art skeleton action recognition models with high confidence. CIASA perturbations show high transferability in black-box settings. We also show that the perturbed skeleton sequences are able to induce adversarial behavior in the RGB videos created with computer graphics. A comprehensive evaluation with NTU and Kinetics data sets ascertains the effectiveness of CIASA for graph-based skeleton action recognition and reveals the imminent threat to the spatiotemporal deep learning tasks in general.**

***Index Terms*— Action recognition, adversarial attack, adversarial examples, adversarial perturbations, skeleton actions, spatiotemporal.**

I. Introduction

SKELETON representation provides the advantage of capturing accurate human pose information while being invariant to action-irrelevant details, such as scene background, clothing patterns, and illumination conditions. This makes skeleton-based action recognition an appealing approach [1]–[6]. The problem is also interesting for multiple application domains, including sport analytics, biomechanics, security, surveillance, animation, and human–computer interactions. Recent contributions in this direction predominantly exploit deep models to encode spatiotemporal dependencies of the skeleton sequences [7]–[10] and achieve remarkable recognition accuracy on benchmark action data sets [11]–[14].

Although deep learning has been successfully applied to many complex problems, it is now known that deep models are vulnerable to adversarial attacks [15], [16]. These attacks can alter model predictions by adding imperceptible perturbations to the input by accessing it at any stage before reaching the model. After the discovery of this intriguing weakness of deep learning [15], many adversarial attacks have surfaced for a variety of vision tasks [17]–[20]. Developing and investigating these attacks not only enhances our understanding of the inner workings of the neural networks [21] but also provides valuable insights for improving the robustness of deep learning in practical adversarial settings.

Deep models for skeleton-based action recognition may also be vulnerable to adversarial attacks. However, adversarial attacks on these models are yet to be explored. A major challenge in this regard is that the skeleton data representation differs significantly from image representation, for which the existing attacks are primarily designed. Human skeleton data are sparse and discrete that evolves over time in rigid spatial configurations. This prevents an attacker from freely modifying the skeletons without raising obvious attack suspicions. Skeleton actions also allow only subtle perturbations along the temporal dimension to preserve the natural action dynamics. In summary, adversarial attacks on skeleton data must carefully account for the skeleton's spatial integrity, temporal coherence, and anthropomorphic plausibility. Otherwise, the attack may be easily detectable. These challenges have so far kept skeleton-based action recognition models away from being scrutinized for adversarial robustness.

This work presents the first adversarial attack on deep skeleton-based action recognition. In particular, we attack the graph convolutional networks (GCNs) [22] for this task [8]. Due to the graph-structured nature of human skeleton, these networks are particularly suitable for processing skeleton data, compared to CNNs that are more amenable to grid-structured inputs such as images. Actions are represented as spatiotemporal graphs that encode intrabody and interframe connections as edges and body joints as nodes. Graph convolution operations are leveraged to model the spatiotemporal dependencies within the skeleton sequences. The physical significance of nodes and edges in these models imposes unique constraints over the potential attacks. For instance, the graph nodes for a skeleton sequence cannot be added or removed because the number of joints in the skeleton must remain fixed. Similarly, the lengths of intrabody edges in the graph cannot be altered arbitrarily as they represent bones. Moreover, interframe edges must always

Manuscript received September 16, 2019; revised April 3, 2020 and August 21, 2020; accepted November 24, 2020. This research was supported by ARC Discovery Grant DP190102443. The GPUs used for this research are donated by the NVIDIA corporation. *(Corresponding author: Naveed Akhtar.)*

The authors are with the Department of Computer Science and Software Engineering, The University of Western Australia, Crawley, WA 6009, Australia (e-mail: jian.liu@research.uwa.edu.au; naveed.akhtar@uwa.edu.au; ajmal.mian@uwa.edu.au).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2020.3043002.

Digital Object Identifier 10.1109/TNNLS.2020.3043002

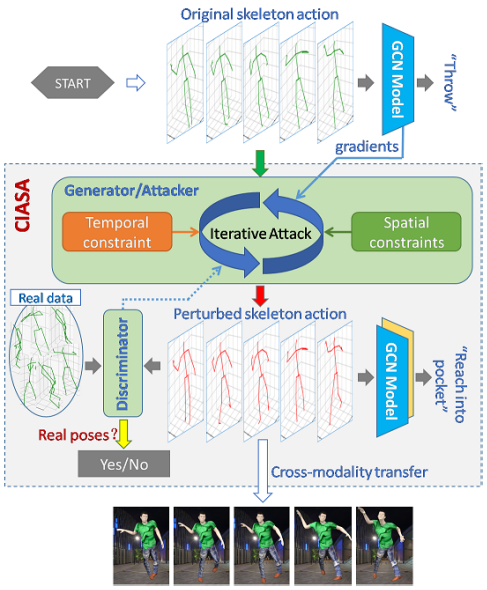


Fig. 1. CIASA schematics. Model gradients are computed for input action sequence to iteratively minimize the model's loss for a target label in small step while accounting for the relevant spatiotemporal constraints. A generator–discriminator framework further ensures anthropomorphic plausibility of the skeletons. Besides cross-model transferability, the attack can also affect RGB videos generated with computer graphics using the skeletons perturbed by CIASA.

connect the same joints along the temporal dimension. Rooted in the skeleton data, such constraints thoroughly distinguish the adversarial attacks on skeleton-based action recognition models from the attacks developed for other kinds of graph networks [23].

We develop an iterative scheme called constrained iterative attack for skeleton actions (CIASA), to generate the desired adversarial skeleton sequences (see Fig. 1). For a given action, CIASA iteratively perturbs its skeleton sequence in small steps to minimize the model prediction loss for a preselected target class while satisfying multiple physical constraints to keep the resulting adversarial sequence natural. In particular, it accounts for spatiotemporal constraints that preserve intraskeleton joint connections, interframe joint connections, and the skeleton bone lengths using a mechanism termed "spatial skeleton realignment" (SSR). For perturbation imperceptibility, it restricts the ℓ_∞-norm of the added noise. In addition, it imposes external temporal dynamics constraints for imperceptible evolution of the adversarial patterns in the skeleton sequence. To further ensure anthropomorphic plausibility of the adversarial skeleton sequence, it exploits the generative adversarial network (GAN) framework [24]. The used GAN configuration reduces the difference between the distribution of adversarial samples generated by our iterative scheme and the clean ground-truth samples. We note that our use of GAN under this setup is an explicit contribution of this article. The proposed GAN setup is generic in its nature that can also be applied to refine perturbations in other application domains.

We analyze the proposed attack by allowing different modes in which CIASA can be used by an attacker. Analogous to standard image-based attacks, we allow perturbation of all skeleton joints in the basic mode. In a localized mode, we provide the flexibility of perturbing only localized regions, e.g., legs of skeleton. This type of attack is particularly suitable to skeleton actions where an attacker may independently alter motion of the least relevant joints for an action to change the prediction. We also introduce an advanced attack mode that further allows a hierarchical magnitude variation in joint perturbations based on the graph structure of the joints.

The notion of localized perturbation also leads to semantically imperceptible perturbations under CIASA where significant perturbation still remains hard to perceive as it is applied to the least significant joints for the original action semantics. Besides demonstrating high fooling rates for the SOTA graph skeleton action recognition model graph convolutional network as a spatial–temporal model (ST-GCN) [8] on NTU [11], [25] and Kinetics [26] data sets, we also show high cross-model transferability of the proposed attack. In addition, we show that videos generated (using computer graphics) from the adversarial skeletons (CIASA's advanced mode) result in lower action recognition accuracy, implying that the attack can be launched in the real world. To the best of our knowledge, this is the first of its kind demonstration of transferability of adversarial attacks beyond a single data modality. In human action recognition, skeletons can sometimes be considered an intermediate product, e.g., when an action recognition system extracts skeleton using a pose estimator. In a typical adversarial setup for deep learning, an adversary has access to the model's input, irrespective of the input being a direct or intermediate product of the sensor [27]. This makes the proposed attack as practical as any existing adversarial attack on deep learning in the digital domain.

II. Related Work

A. Skeleton-Based Action Recognition

The use of skeleton data in action recognition becomes popular as reliable skeleton data can be obtained from modern RGB-D sensors (e.g., Microsoft Kinect) or extracted from images taken from a single RGB camera [28]. A skeleton action is represented as a sequence of human skeletons, which encodes rich spatiotemporal information regarding human motions. Early research in skeleton-based action recognition formulated skeleton joints and their temporal variations as trajectories [2]. Huang *et al.* [29] incorporated the Lie group structure into the task and transformed the high-dimensional Lie group trajectory into temporally aligned Lie group features for skeleton-based action recognition.

To leverage the power of convolutional neural network, Du *et al.* [3] represented a skeleton sequence as a matrix by concatenating the joint coordinates. The matrix is arranged as an image, which can be fed into CNN for recognition. Similarly, Ke *et al.* [5] transformed a skeleton sequence into three clips of grayscale images that encode spatial dependencies between the joints by inserting reference joints. To fit the target neural networks, these methods resize the transformed images. Liu *et al.* [30] proposed a universal unit "skepxel" to create images of arbitrary dimensions for CNN processing. In addition to CNNs, recurrent neural networks are also

employed to model temporal dependencies in skeleton-based human action analysis [31]–[33]. To directly process the sparse skeleton data with neural networks, GCN [22] is used for action recognition. Since GCN is particularly relevant to this work, we review its relevant literature and application to action recognition in more detail.

B. Graph Convolution Networks

The topology of human skeleton joints is a typical graph structure, where the joints and bones are, respectively, interpreted as graph nodes and edges. This makes GCNs particularly more suitable to process skeletons compared with CNNs. Consequently, there have been several recent attempts in modeling human skeleton actions using graph representation and exploiting the spatiotemporal dependencies in skeleton sequences with the help of GCNs.

Yan *et al*. [8] used ST-GCN that aims to capture embedded patterns in the spatial configuration of skeleton joints and their temporal dynamics simultaneously. Along the skeleton sequence, they defined a graph convolution operation, where the input is the joint coordinate vectors on the graph nodes. The convolution kernel samples the neighboring joints within the skeleton frame as well as the temporally connected joints at a defined temporal range.

Tang *et al*. [34] incorporated deep reinforcement learning with graph neural network to recognize skeleton-based actions. Their model distills the most informative skeleton frames and discards the ambiguous ones. As opposed to previous works where joints dependence is limited in the real physical connection (intrinsic dependence), they proposed extrinsic joint dependence, which exploits the relationship between joints that have physical disconnection. Since graph representation of skeleton is crucial to graph convolution, Gao *et al*. [35] formulated the skeleton graph representation as an optimization problem and proposed graph regression to statistically learn the underlying graph from multiple observations. The learned sparse graph pattern links both physical and nonphysical edges of skeleton joints, along with the spatiotemporal dimension of the skeleton action sequences.

To justify the importance of bone motions in skeleton action recognition, Zhang *et al*. [36] focused on skeleton bones and extended the graph convolution from graph nodes to graph edges. Their graph edge convolution defines a receptive field of edges, which consists of a center edge and its spatiotemporal neighbors. By combining the graph edge and node convolutions, they proposed a two-stream graph network, which achieved remarkable performances on benchmark data sets. Similarly, Shi *et al*. [37] also proposed a two-stream framework to model joints and bones information simultaneously.

Wen *et al*. [38] investigated the spatial structure of skeleton joints that are not physically connected so that the semantic roles of the disconnected joints can be exploited. Peng *et al*. [39] took a further step to automatically design graph structure for skeleton-based action through neural architecture search. Other efforts in [38]–[41] focus on exploring the intrinsic high-order relationships within the skeletons, trying to expand the traditional recognition schemes that are based on the fixed joint connectivity. However, the used spatiotemporal blocks remain variable in these cases, and the mix of semantic and nonsemantic connections can cause generalization issues for the networks.

C. Adversarial Attacks on Graph Data

Adversarial attacks [15] have recently attracted significant research attention [27], resulting in few attacks on graph data as well. However, compared with the adversarial attacks for image data [16], [42]–[44], several new challenges appear in attacking graph data [45]. First, the graph structure and features of graph nodes are in discrete domain with certain predefined structures, which leaves a lower degree of freedom for creating adversarial perturbations. Second, the imperceptibility of adversarial perturbations in graph data is neither easy to define nor straightforward to achieve, as the discrete graph data inherently prevent infinitesimal small changes [23].

Dai *et al*. [46] focused on attacking structural information, i.e., adding/deleting graph edges, to launch adversarial attacks on graph-structured data. Given the gradient information of target classifier, one of their proposed attacks modifies the graph edges that are most likely to change the objective. In addition to modifying graph edges, Zügner *et al*. [23] adopted an attack strategy to modify the graph node features as well as graph edge structure. To ensure the imperceptibility of adversarial perturbations, they designed constraints based on power law [47] to preserve the degree distribution of graph structures and feature statistics.

Being atypical graph data, human skeletons have several unique properties. In a human skeleton, the graph edges represent rigid human bones, which connects a finite number of human joints to form a standard spatial configuration. Unlike graph data with mutable graph structure (e.g., social network graph [48]), the human bones are fixed in terms of both joint connections and bone lengths. This property implies that attacking human skeletons by adding or deleting bones will be detected easily by observers. The hierarchical nature of human skeleton data is also different from normal graph data, as in human skeleton the motion of children joints/bones is affected by their parents' behaviors. This chain-like motion kinetics of human skeletons must be considered when launching adversarial attacks on skeleton actions. Hence, despite the existence of adversarial attacks on graph data, robustness of skeleton-based human action recognition against adversarial attacks remains largely unexplored.

In this work, we specifically focus on adversarial attacks on human skeleton sequences to fool skeleton-based action recognition models. To design effective and meaningful attacks, we take the spatial and temporal attributes of skeleton data into account while creating the adversarial perturbations. Due to its widespread use in graph convolution network-based action recognition, we select ST-GCN [8] as our target model and launch our attack against it. However, our attack is generic for similar graph-based model. In the section to follow, we formulate our problem in the context of skeleton-based human action recognition.

III. Problem Formulation

To formulate the problem, we first briefly revisit the ST-GCN [8] for skeleton-based action recognition. Using this prerequisite knowledge, we subsequently formalize our problem of adversarial attacks on skeleton action recognition.

A. Revisiting ST-GCN

An action in skeleton domain is represented as a sequence of T skeleton frames, where every skeleton consists of N body joints. Given such $N \times T$ volumes of joints, an undirected spatiotemporal graph $G = (V, E)$ can be constructed, where V denotes the node set of graph and E is the edge set. Here, $V = \{v_{ti} | t = 1, \ldots, T, i = 1, \ldots, N\}$ encodes the skeleton joints. An element "v" of this set can also be considered to encode a joint's Cartesian coordinates. Two kinds of graph edges E are defined for joints, namely; intrabody edge E^S and interframe edge E^F. Specifically, E^S is represented as an $N \times N$ adjacency matrix of graph nodes, where the matrix element $E^S_{ij} = 1 | i \neq j$ identifies that a physical bone connection exists between the body joints v_i and v_j. The interframe edges E^F denotes the connections of the same joints between the consecutive frames, which can also be treated as temporal trajectories of the skeleton joints.

Given the spatiotemporal skeleton graph G, a graph convolution operation is defined by extending the conventional image-based convolution. Along the spatial dimension, graph convolution is conducted on a graph node v_i around its neighboring nodes $v_j \in \psi(v_i)$

$$f_{\text{out}}(v_i) = \sum_{v_j \in \psi(v_i)} \frac{1}{Z_i(v_j)} f_{\text{in}}(v_j) \cdot w(l_i(v_j)) \tag{1}$$

where ψ is the sampling function to define a neighboring node set for the joint v_i, f_{in} is the input feature map, w is the weight function that indexes convolution weight vectors based on the labels of neighboring nodes v_j, and $Z_i(v_j)$ is the number of neighboring nodes to normalize the inner product. The labels of neighboring nodes are assigned with a labeling function $l : \psi(v_i) \rightarrow \{0, \ldots, K-1\}$, where K defines the spatial kernel size. ST-GCN employs different skeleton partitioning strategies for the labeling purpose. To conduct graph convolution in spatiotemporal dimensions, the sampling function $\psi(v)$ and the labeling function $l(v)$ are extended to cover a predefined temporal range Γ, which decides the temporal kernel size.

ST-GCN [8] adopts the implementation of graph convolution network in [22] to create a nine-layer neural network with temporal kernel size $\Gamma = 9$ for each layer. Starting from 64, the number of channels is doubled for every three layers. The resulting tensor is pooled at the last layer to produce a feature vector $f_{\text{final}} \in \mathbb{R}^{256}$, which is fed to a softmax classifier for predicting the action label. The network mapping function is compactly represented as

$$Z_{G,c} = \mathcal{F}_\theta(V, E) = \arg\max\langle \text{softmax}(f_{\text{final}}) \rangle \tag{2}$$

where θ denotes the network parameters. We use $Z_{G,c}$ to denote the probability of assigning spatiotemporal skeleton graph G to class $c \in C = \{1, 2, \ldots, c_k\}$. After training, the network parameters are fine-tuned to minimize the cross-entropy loss between the predicted class c and the ground truth c_{gt} that maximizes the probability $Z_{G,c} | c = c_{\text{gt}}$ for the data set under consideration.

B. Adversarial Attack on Skeleton Action Recognition

Given an original spatiotemporal skeleton graph $G^0 = (V^0, E^0)$ and a trained ST-GCN model $\mathcal{F}_\theta$, our goal is to apply adversarial perturbation to the graph G^0, resulting in a perturbed graph $G' = (V', E')$ that satisfies the following broad constraint:

$$Z_{G',c} = \mathcal{F}_\theta(V', E'), \quad \text{s.t. } c \neq c_{\text{gt}}. \tag{3}$$

In the following, we examine this objective from various aspects to compute effective adversarial perturbations for the skeleton action recognition.

1) Feature Perturbations: As explained in Section III-A, V denotes the skeleton joints whose elements can be represented as the Cartesian coordinates of joints, e.g., $v_{ti} : \{x_{ti}, y_{ti}, z_{ti}\}$. For a particular node v_{ti} in the skeleton graph G, an adversarial attack can change its original location such that $v'_{ti} = v^0_{ti} + \rho_{ti}$, where $\rho_{ti} \in \mathbb{R}^3$ is the adversarial perturbation for the node v_{ti}. We refer to this type of perturbation as feature perturbation. Another possibility could be to alter the adjacency relationship in a graph such that $E'_{ij} \neq E^0_{ij} | i, j \in \mathcal{V}$, where $\mathcal{V}$ denotes the set of affected graph nodes. However, in a spatiotemporal skeleton graph G, perturbing edges have strong physical implications. Recall that intrabody connections of joints define the rigid bones within a skeleton, and interframe connections define the temporal movements of the joints. Hence, changes to these connections can lead to skeleton sequences that cannot be interpreted as any meaningful human action. Therefore, the objective in (3) must further be constrained to preserve the graph structure while computing the perturbation. To account for that, we must modify the overall constraint to

$$Z_{G',c} = \mathcal{F}_\theta(V', E^0), \quad \text{s.t. } c \neq c_{\text{gt}}. \tag{4}$$

2) Perturbation Imperceptibility: Imperceptibility is an important attribute of adversarial attacks, as adversaries are likely to fool deep models in unnoticeable ways. Here, we explore perturbation imperceptibility in the context of skeleton actions. This leads to further constraints that must be satisfied when launching adversarial attacks on a skeleton graph G.

For the conventional image data, imperceptibility of perturbations is typically achieved by restricting $\|\rho\|_p < \xi$, where $\|\cdot\|_p$ denotes the ℓ_p-norm of a vector with $p \in [0, \infty)$ and ξ is a predefined constant [49]. For the skeleton graph data, however, the graph structure is discrete and graph nodes are dependent on each other, which makes it more challenging to keep a valid perturbation fully imperceptible. We tackle the challenge of perceptibility for skeleton perturbations from multiple points of view that result in multiple constraints for the overall objective, as explained in the following paragraphs.

a) Joints variation constraint: Focusing on feature perturbation on skeleton graph, the location of a target skeleton joint is changed such that $v'_{ti} = v^0_{ti} + \rho_{ti}$. It is intuitive to constrain ρ of every target joint in a small range to

avoid breaking the spatial integrity of the skeleton. Hence, we employ the following constraint:

$$\|\rho_{ti}\|_\infty \leqslant \epsilon_i \mid t \in [1, \ldots, T]; \quad i \in [1, \ldots, N] \tag{5}$$

where $\|\cdot\|_\infty$ denotes ℓ_∞-norm and ϵ_i is a prefixed constant. By restricting the joint variations to a small ℓ_∞-ball, we encourage perturbation imperceptibility. From the implementation view, when the ball radius ϵ is constant for all joints, we call it global clipping of the perturbed joints, and when the value of ϵ_i is joint-dependent, we call it hierarchical clipping.

b) Bone length constraint: In a skeleton graph G, the intrabody graph connections E^S represent rigid human bones, and hence, their lengths must be preserved despite the perturbations. In the case of $E^S_{ij} = 1 | i \neq j$, the length of the bone between joints i and j at frame t can be calculated as $B_{ij,t} = \|v_{ti} - v_{tj}\|_2$. After applying perturbations to the graph, the new bone length $B'_{ij,t} = \|v'_{ti} - v'_{tj}\|_2$ should satisfy the following:

$$B_{ij,t} = B'_{ij,t} \mid t \in [1, \ldots, T] \quad \text{s.t. } E^S_{ij} = 1. \tag{6}$$

c) Temporal dynamics constraint: Due to the spatiotemporal nature of skeleton action graphs, we disentangle the restrictions over perturbations into spatial and temporal constraints. Previous paragraphs mainly focused on the spatial constraints. Here, we analyze the problem from a temporal perspective.

A skeleton action is a sequence of skeleton frames that transit smoothly along the temporal dimension. A skeleton perturbation may lead to random jitters in the temporal trajectories of the joints and compromise the smooth temporal dynamics of the target skeleton action. To address this problem, we impose an explicit temporal constraint over the perturbations. Inspired by [28], we penalize acceleration of the perturbed joints to enforce temporal stability. Given consecutive perturbed skeleton frames f'_{t-1}, f'_t, and f'_{t+1}, the acceleration is calculated as $\ddot{f}'_t = f'_{t+1} + f'_{t-1} - 2f'_t$. Note that, $f'_t = \{v'_{ti} | i = 1, \ldots, N\}$, where N is the number of perturbed skeleton joints. The calculation of acceleration is conducted on individual joints. We optimize our attacker $\mathcal{A}$ (discussed further next) by including the following temporal smoothness loss in the overall objective:

$$\mathcal{L}_{\text{smooth}}(\mathcal{A}) = \frac{1}{T-1} \sum_{t=2}^{T} \ddot{f}'_t = \frac{1}{T-1} \sum_{t=2}^{T} \sum_{i=1}^{N} \ddot{v}'_{ti} \tag{7}$$

where T denotes the number of time steps considered. In the text to follow, we use $\ddot{f}'_t$ to denote the joint acceleration for notational simplification. Minimization of the loss $\mathcal{L}_{\text{smooth}}$ encourages smoothness in the resulting skeleton sequence by restricting any jitters caused by the perturbation.

3) Anthropomorphic Plausibility: After adversarial perturbation is applied to a skeleton, the resulting skeleton can become anthropomorphically implausible. For instance, the perturbed arms and legs may bend unnaturally or significant self-intersections may occur within the perturbed armature. Such unnatural behavior can easily raise attack suspicions. Therefore, this potential behavior needs to be regularized while computing the perturbations.

Let $\mathcal{P}$ define the distribution of natural skeleton graphs. A sample graph G^0 is drawn from this distribution with probability $\mathcal{P}(G^0)$. We can treat an adversarial skeleton's graph G' to be a sample of another similar distribution $\mathcal{P}'$. The latter distribution should closely resemble the former under the restriction of minimal perturbation of joints and anthropomorphic plausibility of the skeletons. Hence, to obtain effective adversarial skeletons, we aim at reducing the distribution gap between $\mathcal{P}$ and $\mathcal{P}'$. To that end, we employ a GAN [24] to learn appropriate distribution in a data-driven manner.

Specifically, we model a skeleton action "attacker" as a function $\mathcal{A}$ such that $G' = \mathcal{A}(G^0)$. In the common GAN setup, the attacker can be interpreted as a generator of perturbed skeletons (see Fig. 1). We set up a binary classification network as the discriminator $\mathcal{D}$. The discriminator accepts either the natural graph $\tilde{G}$ or the perturbed graph G' as its input and predicts the probability that the input graph came from $\mathcal{P}$. The $\tilde{G}$ and G' are kept "unpaired," implying that $\tilde{G}$ and G^0 are different graphs sampled from the distribution $\mathcal{P}$. To formulate the adversarial learning process, we leverage the least squares objective [50] to train the attacker $\mathcal{A}$ and the discriminator $\mathcal{D}$ using the following loss functions:

$$\mathcal{L}_{\text{adv}}(\mathcal{A}) = \mathbb{E}_{G' \sim \mathcal{P}'}[(\mathcal{D}(G') - 1)^2] \tag{8}$$

$$\mathcal{L}_{\text{adv}}(\mathcal{D}) = \mathbb{E}_{\tilde{G} \sim \mathcal{P}}[(\mathcal{D}(\tilde{G}) - 1)^2] + \mathbb{E}_{G' \sim \mathcal{P}'}[\mathcal{D}(G')^2]. \tag{9}$$

During training, $\mathcal{A}$ and $\mathcal{D}$ are optimized jointly. We discuss the related implementation details in Section IV. In our setup, minimizing $\mathcal{L}_{\text{adv}}$ leads to perturbations that visually appear natural to humans by maintaining anthorpomorphic plausibility of the skeletons.

4) Localized Joint Perturbation: Unlike the pixel space of images, a skeleton action graph has highly discrete structure along both spatial and temporal dimensions. This discreteness poses unconventional challenges for adversarial attacks in this domain. Nevertheless, it also gives rise to interesting investigation directions. For instance, it is intriguing to devise a localized adversarial attack, which fools the model by perturbing only a particular part of the skeleton graph. If we closely observe a skeleton action, it is clear that different body joints contribute differently to our perception of actions. In addition, most of the human actions are recognizable by the motion patterns associated with the dominant body parts, e.g., arms and legs. Such observations make localized perturbations particularly relevant to the skeleton data.

Localized joint perturbations allow for less variations in the overall skeleton, which is beneficial for imperceptibility. They also provide a controlled injection of regional modification to the target skeleton action. To allow that, we define a subset of joints within a skeleton as the attack region. Only the joints in that region are modified for localized perturbations. Consequently, all the constraints in Section III-B2 still hold for the attack.

IV. Attacker Implementation

A. One-Step Attack

First, we adopt the fast gradient sign method (FGSM) [16] as a primitive attack to create skeleton perturbation V' in a

single step. This adoption allows us to put our attack in a better context for the active community in the direction of adversarial attacks. For the FGSM-based attack in our setup, the perturbation computation can be expressed as

$$V' = V^0 + \epsilon \ \text{sign}(\nabla_{V^0}\mathcal{L}(\mathcal{F}_\theta(V^0, E^0), c_{\text{gt}})) \tag{10}$$

where $\mathcal{F}_\theta$ denotes the trained ST-GCN [8] model, $\mathcal{L}$ is the cross-entropy loss for action recognition, and ∇_{V^0} is the derivative operation that computes the gradient of ST-GCN loss w.r.t. V^0, given the current model parameters θ and the ground-truth action label c_{gt}. The sign of gradient is scaled with a parameter ϵ and added to the original graph V^0. The FGSM-based attack is computationally efficient as it takes only a single step in the direction of increasing the recognition loss of the target model.

The basic FGSM attack does not specify the label for the misclassified action and therefore is a "nontargeted" attack. If we specify a particular label for c_{gt} in (10) and subtract the gradient's sign from the original graph V^0 [instead of adding it, as in (10)], the resulting attack becomes a targeted attack [51] that is likely to change the predicted label of the considered action to a prespecified label.

B. Iterative Attack

The FGSM attack takes a single step over the model cost surface to increase the loss for the given input. An intuitive extension of this notion is to iteratively take multiple steps while adjusting the step direction [52]. For the iterative attack, we also adopt the same technique for the skeleton graph input. However, here, we focus on targeted attacks. This is because targeted attacks are more interesting for real-world applications and nontargeted attacks can essentially be considered a degenerate case of the targeted attack, where the target label is chosen at random. Hence, an effective targeted attack already ensures nontargeted model fooling. To implement, we specify the desired target class and take multiple steps in the direction of decreasing the prediction loss of the model for the target class.

We implement the iterative targeted attack while enforcing the constraints discussed in Section III-B2. The resulting algorithm is termed CIASA. At the core of CIASA is the following iterative process:

$$V'_0 = V^0; \ \ V'_{N+1} = \mathcal{C}\big(V'_N - \alpha \ \big(\nabla_{V'_N}\mathcal{L}_{\text{CIASA}}\big(V'_N, c_{\text{target}}\big)\big)\big). \tag{11}$$

At each iteration, V'_N is adjusted toward the direction of minimizing the overall CIASA loss $\mathcal{L}_{\text{CIASA}}$ using a step size α. This is equivalent to a gradient descent iteration with α as the learning rate, where the skeleton graph V'_N is treated as the model parameter. Hence, we directly exploit the Adam optimizer [53] in the PyTorch library[1] for this computation. The operation $\mathcal{C}(.)$ in (11) truncates and realigns the values in its argument with preset conditions, explained next.

In (11), the overall CIASA loss $\mathcal{L}_{\text{CIASA}}$ consists of the following components:

$$\mathcal{L}_{\text{CIASA}} = \mathcal{L}_{\text{pred}} + \lambda(\mathcal{L}_{\text{smooth}} + \mathcal{L}_{\text{adv}}) \tag{12}$$

[1]https://pytorch.org/

Algorithm 1 Constrained Iterative Attacker $\mathcal{A}$ to Fool Skeleton-Base Action Recognition

Input: Original graph nodes $V^0 \in \mathbb{R}^{3\times N\times T}$, trained ST-GCN model $\mathcal{F}_\theta()$, desired target class c_{target}, perturbation clipping factor ϵ, max_iter=M, learning rate α
Output: Perturbed graph nodes $V' \in \mathbb{R}^{3\times N\times T}$.

1: set initial $V' = V^0$
2: **while** $i < M$ **do**
3: feed forward $Z = \mathcal{F}_\theta(V')$
4: $\mathcal{L}_{pred} = \text{CrossEntroyLoss}(Z, c_{target})$
5: $\mathcal{L}_{smooth} = \frac{1}{T-1}\sum_{t=2}^{T} \ddot{f}'_t$
6: $\mathcal{L}_{adv}(\mathcal{A}) = (\mathcal{D}_\omega(\mathcal{A}(V')) - 1)^2$
7: $\mathcal{L}_{adv}(\mathcal{D}) = (\mathcal{D}_\omega(\tilde{V}) - 1)^2 + \mathcal{D}_\omega(V')^2$
8: $\mathcal{L}_{\text{CIASA}} = \mathcal{L}_{pred} + \lambda(\mathcal{L}_{smooth} + \mathcal{L}_{adv})$
9: $(\mathcal{L}_{\text{CIASA}}).\text{Backward}() \Rightarrow gradients$
10: $V', \omega = \text{AdamOptimizer}([V', \omega], \ gradients)$
11: **if** $|V' - V^0| > \epsilon$ **then**
12: $V' = \text{Clip}(V') \sim [V^0 - \epsilon, V^0 + \epsilon])$
13: **end if**
14: Skeleton realignment $V' = \text{SSR}(\text{V}')$
15: $i = i + 1$
16: **end while**
17: **return** V'

where $\mathcal{L}_{\text{pred}}$ is the cross-entropy loss of the model prediction on V' for the desired target class c_{target} and $\mathcal{L}_{\text{smooth}}$ is the temporal smoothness loss calculated according to (7). GAN regularization loss $\mathcal{L}_{\text{adv}}$ is a combination of $\mathcal{L}_{\text{adv}}(\mathcal{A})$ and $\mathcal{L}_{\text{adv}}(\mathcal{D})$ given in (8) and (9). λ is a weighting hyperparameter to balance the individual loss components.

Implementing the process identified by (11) produces the perturbed skeleton V' that fools the model into misclassifying the original action as c_{target} while complying to the spatiotemporal constraints derived in Section III. The pseudocode of implementing the process of (11) as CIASA is presented in Algorithm 1. The algorithm starts with a forward pass of V' through the target model $\mathcal{F}_\theta()$, i.e., ST-GCN. The respective losses are then computed to form the overall CIASA loss $\mathcal{L}_{\text{CIASA}}$. At line 6 and 7, $\mathcal{L}_{\text{adv}}$ is computed according to the losses defined in (8) and (9). Here, we replace G with V based on the algorithm context. $\mathcal{D}_\omega$ denotes the discriminator network that is parameterized by ω. Note that, the real data $\tilde{V}$ and the perturbed data V' are unpaired, as discussed in Section III-B3. On line 9, the gradient information is obtained through the backpropagation operation denoted as ".Backward()." We employ the Adam optimizer [53] to update the skeleton joints V' and the discriminator parameters ω. Clipping operation is then applied to truncate V' to preset ranges. In our case, the scaling factor ϵ restricts the ℓ_∞-norm of the perturbation at graph nodes. For global clipping, $\epsilon \in \mathbb{R}$ is a scalar value that results in equal clipping on all joints. For the hierarchical clipping, $\epsilon \in \mathbb{R}^N$ defines different clipping strengths for different joints. The clipping imposes the joint variation constraint over the perturbations. To impose the bone length constraint, SSR is proposed to realign the skeleton bones within the clipped V' according to the original bone

lengths. Note that the operations of clipping and realignment constitute the function $\mathcal{C}(.)$ shown in (11). We empirically set the weight factor λ as 10 and the base learning rate for the Adam optimizer α as 0.01. In the following, we discuss the implementation of SSR and discriminator network $\mathcal{D}$.

1) Spatial Skeleton Realignment: We propose SSR to preserve the bone length constraint as we perturb the skeleton graph. SSR is executed at each iteration after V' is updated and clipped in order to realign every perturbed skeleton frame based on the original bone lengths. Specifically, for every updated skeleton joint v'_j, we find its parent joint v'_i along the intrabody edge E^S. The bone between joints i and j is defined as a vector $b'_{ij} = v'_j - v'_i$. Then, we modify the joint v'_j along the vector direction $\overline{b'_{ij}}$ to meet the constraint in 6. The modification applied to v'_j is also applied to all of its children/grandchildren joints. To complete the SSR, the above process starts from the root joint and propagates through the whole skeleton.

2) GAN Regularization: To enforce the anthropomorphic plausibility of the perturbed skeleton action, the adversarial regularization term $\mathcal{L}_{\text{adv}}$ is optimized jointly with the other attack objectives. Taking per-frame skeleton feature map, say X as the input, a discriminator network $\mathcal{D}$ is trained to classify the skeleton as *fake* or *real* (i.e., perturbed versus original), while the attacker $\mathcal{A}$ is competing with $\mathcal{D}$ to increase the probability of the perturbed skeleton being classified as real.

We leverage the angles between skeleton bones to construct the feature map X. For a pair of bones b_{ij} and b_{uv}, the corresponding element in the feature map is defined as the cosine distance between the bones as

$$x_{ij-uv} = \frac{b_{ij} \cdot b_{uv}}{\|b_{ij}\| \|b_{uv}\|}. \tag{13}$$

We select a group of major bones to construct the feature map X, while insignificant bones of fingers and toes are excluded to avoid unnecessary noise. The resulting feature map has dimension $X \in \mathbb{R}^{C,H,W}$, where $C = 1$ and $H = W$ equals the number of selected bones. We model $\mathcal{D}$ as a binary classification network that consists of two convolution layers and one fully connected layer. The convolution kernel size is 3, and the number of channels produced by the convolution is 32. $\mathcal{D}$ outputs values in the range [0, 1], signifying the probability that X is a real sample.

V. Experiments

In the following, we evaluate the effectiveness of the proposed attack for skeleton-based action recognition. We examine different attack modes on standard skeleton action data sets. We also demonstrate the transferability of attack and explore generalization of the computed adversarial perturbations beyond the skeleton data modality. Finally, an ablation study is provided to highlight the contributions of various constraints to the overall fooling rate achieved by the proposed attack.

A. Data Set and Evaluation Metric

1) NTU RGB + D [11]: NTU RGB + D Human Activity data set is collected with Kinect v2 camera and includes 56 880 action samples. Each action has RGB, depth, skeleton, and infrared data associated with it. However, we are only concerned with the skeleton data in this work. For the skeleton-based action recognition with ST-GCN, we follow the standard protocols defined in [11], i.e., cross-subject and cross-view recognition. Accordingly, two different ST-GCN models are used in our experiments, one for each protocol. We denote these models as NTU_{XS} and NTU_{XV} for cross-subject and cross-view recognition. While the original data set is split into training and testing sets, we only manipulate the testing set, as no separate training data are required for the attack.

2) NTU RGB + D 120 [25]: This is an extension of the NTU RGB + D data set. Compared to the original data set, the NTU RGB + D 120 extends the number of action types from 60 to 120 and includes more subjects, backgrounds, and camera viewpoints. In total, 114 480 RGB + D video samples are captured for 106 distinct human subjects. The evaluation protocol is defined as cross subject and cross setup for this data set, and we denote them as NTU_{XS} and NTU_{XT}, respectively, in our experiments.

3) Kinetics [26]: Kinetics data set is a large unconstrained action data set with 400 action classes. For skeleton-based action recognition using this data, the original ST-GCN [8] first uses OpenPose [54], [55] to estimate 2-D skeletons with 18 body joints. Then, the estimation confidence "c" for every joint is concatenated to its 2-D coordinates (x, y) to form a tuple (x, y, c). The tuples for all joints in a skeleton are collectively considered as an input sample by the ST-GCN model. For the adversarial attack, we mask the channel of confidence values and only perturb the (x, y) components for the Kinetics data set.

4) Evaluation Metric: The evaluation metric used to evaluate the success of adversarial attacks is known as fooling rate [27]. It indicates the percentage of data samples over which the model changes its predicted label after the samples have been adversarially perturbed. In the adversarial attacks literature, this is the most commonly used metric to evaluate an attack's performance [27]. In the case of targeted attacks, it determines the percentage of the samples successfully misclassified as the target label after the attack.

B. Nontargeted Attack

As it is the first work in the direction of attacking skeleton-based action recognition, it is important to put our attacking technique into perspective. Hence, we first conduct a simpler nontargeted attack on the NTU RGB + D and Kinetics data sets using the one-step attack discussed in Section IV-A [see (10)]. We compute the fooling rates for both data sets under different values of the perturbation scaling factor ϵ. Both cross-view and cross-subject protocols were considered in this experiment for the NTU RGB + D data set. The fooling rates achieved with the one-step method for various ϵ values are summarized in Fig. 2. As can be seen, the nontargeted fooling is reasonably successful under the proposed formulation of the problem for skeleton-based action recognition. The fooling rates for all protocols remain higher than 90% once the ϵ value reaches 0.02. This is still a reasonably small perturbation value

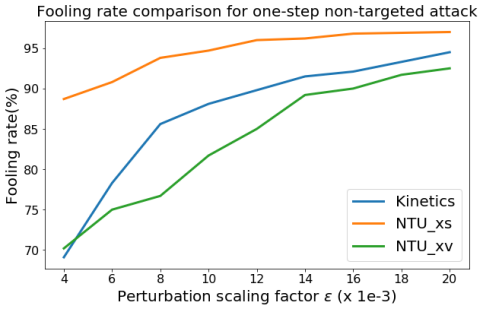


Fig. 2. Fooling rates (%) achieved by one-step nontargeted attack with different perturbation scaling factors for NTU RGB + D and Kinetics data sets. Both cross-subject NTU_{XS} and cross-view NTU_{XV} protocols are considered for the NTU RGB + D data set.

that is equivalent to one twentieth of the average skeleton height.

To visualize perturbed skeletons, Fig. 3(a) shows a successful attack on the NTU RGB + D data set for cross-view fooling. The original and perturbed skeletons are plotted with green and red colors, respectively. Note that, in this illustration and the examples to follow, we provide a positional offset between different skeletons for better visualization. For the shown sequence of skeleton frames, the original label is "Brush hair," that is predicted as "Wipe face" after the attack is performed. The temporal dimension evolves from left to right. Ignoring the positional offset, it is easy to see that the perturbation generally remains hard to perceive in the skeleton.

C. Targeted Attack

We use the proposed CIASA attacker explained in Section IV-B (see Algorithm 1) to conduct targeted attacks on both NTU and Kinetics data sets. We specify the least-likely action prediction of the ST-GCN models as the target label c_{target} as described in (11), implying that the most challenging misclassification target is chosen to launch attacks. CIASA is configured to launch attacks in three modes; namely, basic mode, localized mode, and advanced mode. In the following, we discuss these modes along the experimental results.

Fig. 3(b) shows an example of CIASA attack in the basic mode. We apply the global clipping discussed in Section III-B2 in this attack mode, where all the skeleton joints are perturbed with the same scaling factor $\epsilon = 0.02$. With this setting, the original action of "Cheer up" in Fig. 3(b) is misinterpreted as "Kicking" with confidence score 99.4%. In the basic mode, the comparison of fooling rates with different ϵ values for the two NTU benchmark data sets are summarized in Table I. First, the results demonstrate successful fooling even for very low ϵ values. Second, it is noteworthy that for similar ϵ values, higher fooling rates are generally achieved by CIASA for targeted fooling compared with the nontargeted fooling of the one step method in Fig. 2. This demonstrates the strength of CIASA as a targeted attack. In our experiments, we observed that the least-likely label of ST-GCN model remains similar for multiple actions. While the presented results do not diversify the target labels of such actions to strictly follow the evaluation protocol, it is possible to manually do so. Loosening the evaluation criterion on these lines will further improve the fooling rate of CIASA.

TABLE I

FOOLING RATES (%) ACHIEVED BY CIASA TARGETED ATTACK (BASIC MODE) WITH DIFFERENT GLOBAL CLIPPING STRENGTH ϵ FOR NTU AND KINETICS DATA SETS. THE NTU RESULTS ARE REPORTED ON THE CLASSIC 60-ACTION NTU DATA SET AND THE LATEST 120-ACTION ENHANCED VERSION, DENOTED AS NTU^{60} AND NTU^{120}, RESPECTIVELY. THE EVALUATION PROTOCOLS FOR NTU DATA SET INCLUDE "CROSS-SUBJECT (XS)," "CROSS-VIEW (XV)," AND "CROSS-SETUP (XT)," MARKED AS SUBSCRIPTS

ϵ ($\times$ 1e-3)	4	6	8	10	12
Kinetics	82.5	92.5	96.5	97.5	99.3
$\text{NTU}^{60}_{\text{XS}}$	89.4	96.6	98.7	99.2	99.8
$\text{NTU}^{60}_{\text{XV}}$	78.2	85.5	93.3	98.9	99.6
$\text{NTU}^{120}_{\text{XS}}$	80.8	87.5	89.5	97.2	99.6
$\text{NTU}^{120}_{\text{XT}}$	72.5	83.3	90.7	95.8	99.2

In Fig. 3(c), we shows an example of CIASA attack in the localized mode, where the localized joint perturbation discussed in Section III-B4 is applied. In this example, two legs of skeleton are set to be the attack regions, which allow eight active joints for perturbations. The remaining joints are unaffected by the computed perturbations. Compared with the basic mode, fewer joints contribute to the overall perturbation in the localized mode. To compensate for the reduced number of active joints, we loose the perturbation scaling factor and set ϵ to 0.08 for this experiment. For the shown example, CIASA achieves fooling with 93.2% confidence for this mode, which is still competitive to the 99.4% confidence in the basic mode.

To further evaluate the localized mode of CIASA with different attack regions, we split the skeleton joints into set sets, as shown in Fig. 4. Then, we conduct CIASA localized attack on the NTU RGB + D data set for the four sets separately. Global clipping is applied for these experiments with the scaling factor $\epsilon = 0.04$. The chosen value of ϵ is intentionally kept lower than that in Fig. 3(c) because we focus on analyzing the fooling prowess of different attack regions instead of simply achieving high fooling rates for all the regions. The results of our experiments are summarized in Table II. It is clear that the CIASA localized attack achieves impressive fooling rates by perturbing only a small set of joints within the skeleton. In addition, different sets of active joints affect the fooling performance differently. In Table II, set-1 and set-2 achieve higher fooling rates than the other two sets. This can be explained by the observation that many dominant movements in the NTU RGB + D data set occur at the upper part of human body.

We also extend the localized mode of CIASA to an advanced mode by replacing the global clipping by hierarchical clipping discussed in Section III-B2. In that case, the scalar clipping value ϵ is replaced by $\epsilon \in \mathbb{R}^N$, where N is the number of active joints to be perturbed. Here, we allow various active joints to change in predefined ranges by using differentiated clipping values. One strategy to differentiate the

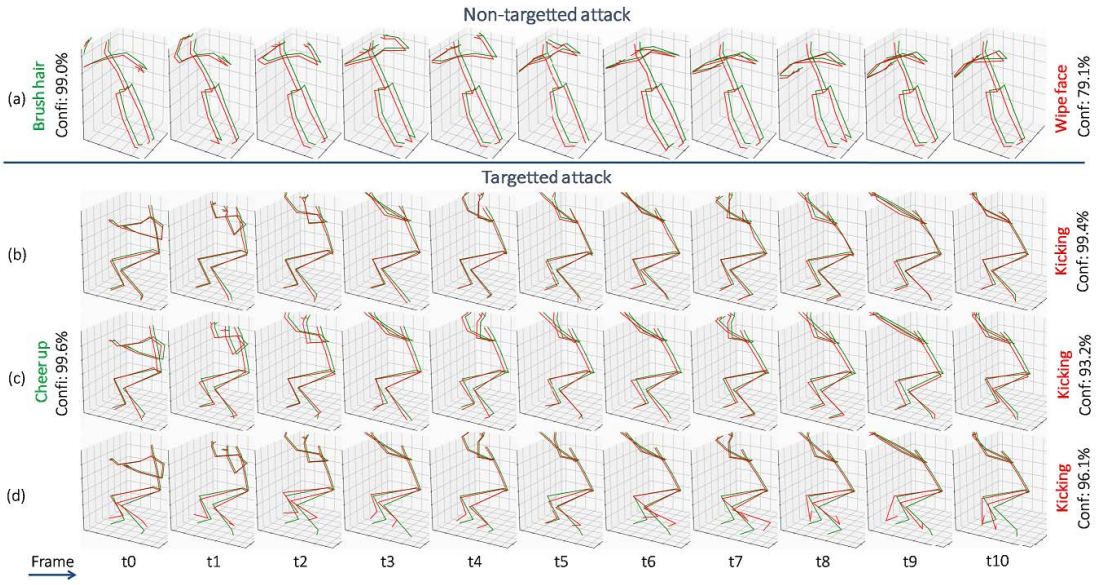


Fig. 3. (a) Top—one-step attack with $\epsilon = 0.02$ is shown where "brush hair" action is misclassified as "wipe face." Bottom: CIASA targeted attack in different modes are shown. (b) Basic mode that perturbs all joints with $\epsilon = 0.01$. (c) Localized mode with only two legs allowed to be perturbed. Global clipping is applied with $\epsilon = 0.08$. (d) Advanced mode where the same two legs are perturbed with hierarchical clipping. The attacks in all modes successfully fool the recognition model with confidences higher than 90%. The temporal dimension evolves from left to right.

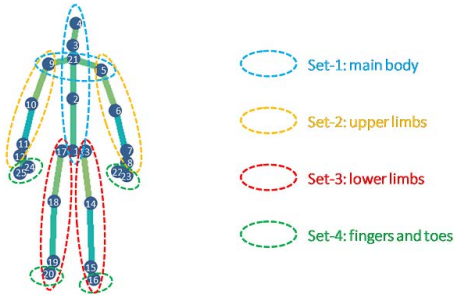


Fig. 4. Skeleton of NTU data set is spitted into four attack regions, each of which is activated to apply CIASA localized attacks. Every attack region consists of roughly the same number of joints.

TABLE II

FOOLING RATE (%) ACHIEVED BY CIASA TARGETED ATTACK (LOCALIZED MODE) WITH DIFFERENT ATTACK REGIONS ON THE NTU RGB + D DATA SET. BOTH CROSS-SUBJECT AND CROSS-VIEW PROTOCOLS ARE EVALUATED. GLOBAL CLIPPING STRENGTH IS SET TO $\epsilon = 0.04$

Attack region	set-1	set-2	set-3	set-4
NTU_{XS}	90.8	93.3	61.3	83.3
NTU_{XV}	85.2	91.7	60.0	81.7

clipping strength is applying incremental ϵ variables from parent joints to their children joints, based on the observation that children joints normally move in larger ranges than their parents. Fig. 3(d) shows an example of successful advanced attack on the NTU RGB + D data set with two legs activated for the attack. The ϵ variables are set to 0.01, 0.05, 0.15, and 0.25 for the joint hips, knees, ankles, and feet, respectively. Note that we intentionally amplify the perturbation ranges at certain joints such as ankles and feet, which results in noticeable perturbations at the attack region. We will justify the intuition behind this differential clipping in the paragraphs to follow. For now, note that in Fig. 3(d), the original label of "Cheer up" is misclassified as "Kicking" with a confidence score 96.1% with the advanced attack.

Although the CIASA attack in *advanced mode* apparently sacrifices the visual imperceptibility of the perturbation, it is able to maintain the "semantic imperceptibility" for the perturbed skeleton. We corroborate this claim with the following observations. First, in Fig. 3(d), the dominant body movements for "Cheer up" action mainly occur in the upper part of the skeleton, whereas the fooling is conducted by perturbing the lower body to which less attention is paid for this action. Consequently, the attack does not incur significant perceptual attention in the first place. Furthermore, due to the spatiotemporal constraints with CIASA attacks, the injected perturbation patterns remain smooth and natural. This further reduces the attack suspicions, compared to any small but unnatural perturbations, e.g., shakiness around the joints.

Further to the above discussion, the perturbations generated in the advanced mode can not only fool the recognition model in skeleton spaces but can also be imitated and reproduced in the physical world. Imagine an end-to-end skeleton-based action recognition system using a monocular camera as its input sensor. For that, RGB images taken from the physical world are first converted to skeleton frames, which are then passed through the skeleton-based action recognition model. For this typical pipeline, it may be inconvenient to interfere with the intermediate skeleton data for the attacking purpose. However, the adversarial perturbations can be injected into the input RGB data by performing an action in front of the camera while imitating the perturbation patterns with selective body parts. The advanced mode of CIASA allows the discovery

of perturbation patterns for such attacks. This is elaborated further in Section V-D2 with relevant context.

D. Transferability of Attack

We examine the transferability of the proposed CIASA attack from two perspectives. First, we evaluate the cross-model transferability of the generated perturbations. Concretely, we attack a skeleton action recognition model A to generate perturbed skeletons. Then, we predict the label of the perturbed skeletons using model B and examine the fooling rate for model B. We chose ST-GCN as model A and separately chose 2s-AGCN [10] and PA-LSTM [11] as model B in our experiments.

Second, we analyze the cross-modality transferability of CIASA attack, i.e., we generate perturbations for one data modality and test their fooling capability in another data modality. We formulate this task as transferring perturbations from skeleton data to RGB data, as RGB cameras are widely used as input sensors for real-world systems. For the cross-modality test, we generate perturbed skeletons by attacking the ST-GCN. Then, those skeletons are converted to RGB actions using a graphics rendering pipeline. To examine whether the adversarial information can be preserved during the conversion, we predict the label of RGB actions under the usual skeleton-based action recognition pipeline for the ST-GCN.

1) Cross-Model Transferability: The 2s-AGCN [10] is a two-stream adaptive GCN for skeleton-based action recognition. This network is significantly different from the ST-GCN [8] as it models a learnable topology of the skeleton graph. In addition to the joint locations, 2s-AGCN also models the bone directions, which results in a two-stream network structure. Different from the GCN-based structure in 2s-AGCN, PA-LSTM [11] employs a recurrent neural network to model the interactions between body parts or with other subjects. While recurrent network architectures are fundamentally different from GCNs, we still explore the transferability of our attack to these networks.

We first generate perturbed skeleton actions based on the ST-GCN model. The basic mode of CIASA with global clipping is employed, where the perturbation scaling factor ϵ is empirically set to 0.012. The cross-view protocol of NTU RGB + D data set is adopted to create perturbed skeletons, which are then evaluated by 2s-AGCN models. We compare the change of recognition accuracy before and after the attack and record the fooling rates for three different configurations of the 2s-AGCN, i.e., joint only (Js-AGCN), bone only (Bs-AGCN), and ensemble (2s-AGCN). We apply the same settings on the PA-LSTM model. The results in Table III show that the perturbations generated with ST-GCN significantly degrade the recognition performance of 2s-AGCN. Moreover, a noticeable performance drop is also observed for PA-LSTM. This demonstrates that the proposed CIASA attack does not only transfer to a different model with a similar architecture type (i.e., Graph-to-graph) but it also transfers reasonably well to different network types (graph-to-recurrent network). This testifies the ability of the proposed attack to fool skeleton-based action recognition models. Broadly, it demonstrates that the proposed CIASA attack is able to generalize well on "unseen" action recognition models.

TABLE III
COMPARISON OF CROSS-MODEL RECOGNITION ACCURACY (%) AND FOOLING RATE (%) ON THREE CONFIGURATIONS OF 2S-AGCN AND PA-LSTM FOR CROSS-VIEW NTU RGB + D PROTOCOL. "ORIGINAL ACCURACY" IS ON CLEAN DATA. "ATTACKED ACCURACY" IS ON PERTURBED DATA

Model	Js-AGCN	Bs-AGCN	2s-AGCN	PA-LSTM
Original Acc.	93.7	93.2	95.1	68.5
Attacked Acc.	13.5	6.8	11.8	52.7
Fooling rate (%)	86.1	93.1	88.4	28.6

2) Cross-Modality Transferability: To transfer the perturbations from skeleton to RGB space, we adopt a human pose synthesis technique [56] to create RGB actions based on the perturbed skeleton sequences generated with the advanced mode of CIASA. The adopted synthesis pipeline can produce realistic RGB actions with diversified human models, backgrounds, cloth textures, and illuminations. Moreover, the temporal dynamics of the underlying action is also reproducible in the synthesized RGB video. We demonstrate successful cross-modality transferability in Fig. 5. Fig. 5(a) and (d) shows the original and perturbed skeleton sequences, respectively. Fig. 5(b) and (e) shows the RGB actions generated using [56] with Fig. 5(a) and (d) used as the inputs skeleton sequences.

First, the successful generation of realistic RGB videos in Fig. 5(b) and (e) affirms that the skeleton perturbations generated by CIASA are useful in producing action perturbations in the physical world beyond the skeleton space. Second, we observe that the adversarial information remains largely preserved during the cross-modality transfer. In Fig. 5, we use VNect [28] as a 3-D pose extractor to recover 3-D skeletons directly from the synthesized RGB actions. The recovered skeleton sequences are then fed to the trained ST-GCN model for action recognition, mimicking the typical pipeline for the skeleton-based action recognition for RGB sensors. We applied the same backgrounds and cloth textures to render videos of both original and perturbed skeletons, to minimize the performance variations of 3-D pose extractor and to keep all other factors fixed. In addition, distinct colors for the backgrounds and the textures are selected to enhance the accuracy of pose extraction. Now, the only difference between the top sequence [see Fig. 5(b)] and bottom sequence [see Fig. 5(e)] is the applied perturbation and it still changes the decision of the network from throw (correct) to back pain (incorrect).

The VNect-recovered 3-D skeletons from clean and perturbed RGB data are, respectively, shown in Fig. 5(c) and (f). As can be seen, the recovered skeletons generally follow the motion patterns encoded in the respective source skeletons. For the clean data, the recovered skeletons in Fig. 5(c) and the source skeletons in Fig. 5(a) are both correctly recognized as "Throw" action. For the perturbed data, the recovered skeleton sequence in Fig. 5(f) has fooled the ST-GCN into

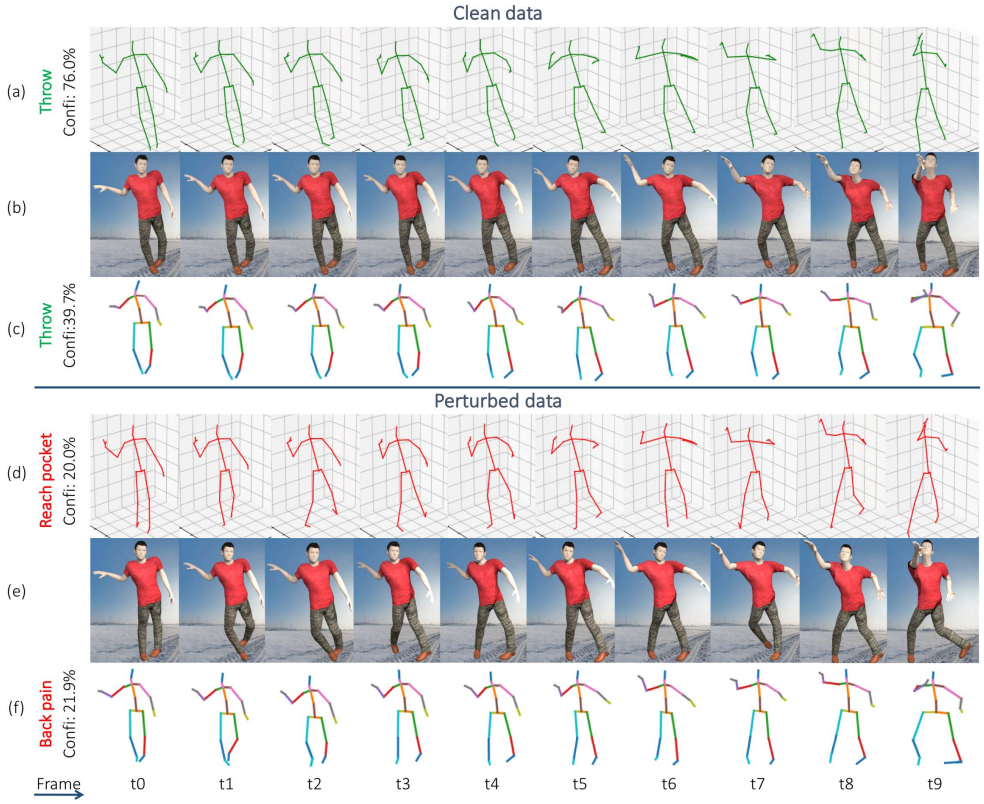


Fig. 5. Top: clean data of different modalities. (a) Original skeleton sequences. (b) RGB video rendered from the original sequence. (c) Recovered 3-D pose sequence extracted from (b) using VNect [28]. Bottom: perturbed data of different modalities. (d) Perturbed skeleton sequences created with the advanced mode of CIASA. (e) RGB video rendered from (d). (f) 3-D poses extracted from (e) using VNect [28]. Note that the same background and texture are used to render the synthetic RGB sequences, to minimize the possible variations imported by the skeleton extractor.

misclassifying the action as "Back pain." Although the fooling is not in the exact least likely class as in Fig. 5(d), misclassification due to CIASA attack for this very challenging scenario is still intriguing. We note that the attack here naturally degenerates into an untargeted attack.

To further scale up the cross-modality experiment, we randomly select 240 skeleton actions for the cross-view protocol of the NTU RGB + D data set. Then, we conduct the cross-modality transfer for all those sequences. We only use a subset of the NTU RGB + D data set because of the high computational time required to render videos for the complete data set. Subsequently, we predict action labels with ST-GCN on the VNect-recovered skeleton sequences for both clean and perturbed data. With this setting, the recognition accuracy is recorded as 54.1% for the clean data and 39.3% for the perturbed data. Compared with the original NTU RGB + D cross-view accuracy of 88.3% [8], lower performance is observed on the clean data due to inaccurate 3-D pose extraction by VNect. Nevertheless, the proposed attack is still able to preserve its adversarial characteristics to further cause a significant accuracy drop in this challenging scenario.

E. Ablation Study

For the CIASA attack, we have proposed a set of spatiotemporal constraints to achieve high-quality adversarial perturbations in terms of both temporal coherence and spatial integrity of the perturbed skeletons. Here, we provide an ablation study to compare the contributions of these constraints in the overall results.

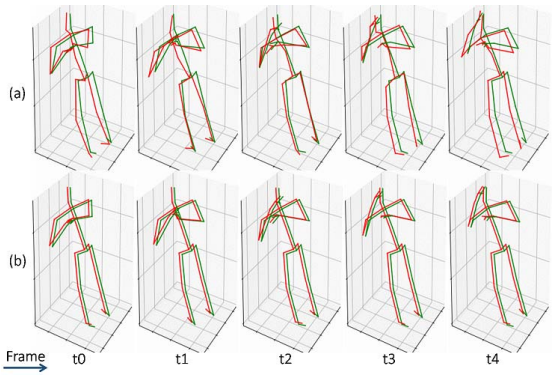


Fig. 6. Temporal smoothness in CIASA. (a) Perturbed skeleton sequence without temporal smoothness constrains. (b) Perturbed sequence with temporal smoothness constrains. The original and perturbed skeletons are shown in green and red colors, respectively.

To enforce temporal smoothness in the perturbed skeleton sequences, we penalize the joint accelerations between the consecutive skeleton frames. Fig. 6 compares the perturbed skeletons with and without this temporal constrains in the basic mode of CIASA, where the original and perturbed skeletons are highlighted with green and red colors, respectively. It is apparent that the perturbed skeletons in Fig. 6(b) move more smoothly than those in Fig. 6(a) along the temporal dimension. This ascertains the effectiveness of temporal smoothing in our

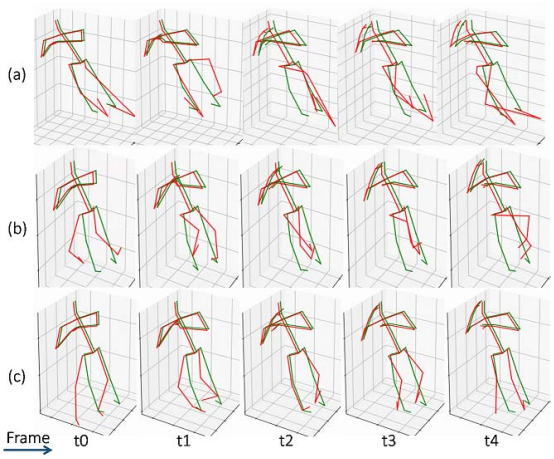


Fig. 7. Effectiveness of spatial constraints in CIASA. A localized attack is launched on two legs of the skeleton. (a) No spatial constraints: pose configuration and bone lengths change randomly. (b) SSR: constrained consistent bone lengths, but unnatural poses. (c) GAN regularization: realistic poses that can correspond to the real-world skeleton motions.

attack. Both perturbations in Fig. 6(a) and (b) successfully fool the recognition model to misclassify "Drink action" as the "Jump up" action.

To enforce spatial integrity and anthropomorphic plausibility, we use SSR and GAN regularization. Such spatial constraints are particularly important for the CIASA localized attacks, where only a given subset of the joints is permitted to be changed. Fig. 7 compares the perturbation results with and without the spatial constraints for a localized attack on skeleton legs. Without any spatial constraints, the perturbed skeletons in Fig. 7(a) shows unrealistic pose configurations and arbitrary lengths of bones. With only SSR enabled in Fig. 7(b), lengths of the perturbed bones are more consistent with their original values; however, the resulting poses are still not realistic in terms of plausibility. By adding the GAN regularization, the skeletons in Fig. 7(c) are more realistic. The skeleton sequences in the figure clearly demonstrate the effectiveness of SSR and GAN regularization in our attack. All sequences in Fig. 7(a)–(c) successfully fool the recognition model in predicting the label "Drink water" as "Jump up."

To quantitatively investigate the contribution of each constraint in CIASA algorithm, we analyze our results for the experiments related to Figs. 6 and 7. A key criterion to evaluate an adversarial attack is to observe the tradeoff between the perceptibility of perturbation and the fooling ratio of attack. Hence, we observe the difference between the perturbed and the original skeleton sequences, given that our perturbation achieves similar fooling rates under different settings of the constraints. Specifically, we launch attacks with global ϵ of $1e^{-}3$ and allow all joints to be perturbed. We report results for four settings in Table IV where we gradually add constraints to our attack. For each setting, we record the fooling rate and the per-skeleton joint variations (min, max, and mean). Table IV shows that given similar (saturated) fooling capacity, the proposed spatial and temporal constraints effectively improve the imperceptibility of the perturbation by reducing the per-skeleton joint variation.

TABLE IV

COMPARISON OF JOINTS VARIATION (MIN, MAX, AND MEAN VALUES) BETWEEN THE ORIGINAL AND PERTURBED SKELETON SEQUENCES UNDER DIFFERENT CONSTRAINTS AND THE RESPECTIVE FOOLING RATES. WE LAUNCH THE BASIC MODE CIASA ATTACK ON NTU CROSS-VIEW DATA SET WITH GLOBAL CLIPPING $\epsilon = 1E^{-}3$. ROW WISE, WE HAVE NO CONSTRAINTS, SSR CONSTRAINT, SSR + TEMPORAL CONSTRAINT, AND SSR + TEMPORAL CONSTRAINT + GAN REGULARIZATION. AN EXAMPLE FIGURE FOR EACH CASE IS MENTIONED IN COLUMN 2

Constraint	Expl.	Min	Max	Mean	Fooling Rate
None	Fig.7a	6.02	9.16	7.59	99.6
SSR	Fig.7b	5.37	9.08	7.24	99.2
SSR+Temp	Fig.6b	5.09	9.05	7.23	99.3
SSR+Temp+GAN	Fig.7c	4.78	9.02	7.09	99.1

TABLE V

EFFECT OF VARYING "λ" IN (12). THE MIN, MAX, AND MEAN VALUES ARE REPORTED FOR JOINT VARIATION BETWEEN THE ORIGINAL AND PERTURBED SKELETONS

λ	Min	Max	Mean	Fooling Rate
5	5.19	9.09	7.73	99.3
10	4.78	9.02	7.09	99.1
20	4.01	8.13	6.45	98.2

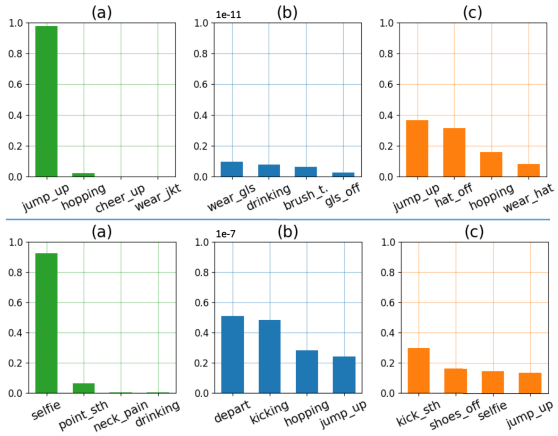


Fig. 8. Typical failure case examples. The advanced mode of CIASA targeted attack is used. (a) Most likely predictions for the original data. (b) Least likely predictions for the original data, where the right-most bar shows the target label of the attack. (c) Most likely predictions for the perturbed data. The first row shows representative examples when the attack failed; however, it significantly reduced the prediction confidence of the correct class. In the second row, the targeted attack degenerates into a nontargeted attack, i.e., successful fooling but to a nontarget class.

For (12), we choose $\lambda = 10$ in our experiments. This hyperparameter also trades off perturbation perceptibility with the fooling rate. Our experiments show that the results of our overall technique remain in acceptable range for $\lambda \in [5, 20]$. In Table V, we report the results in this range. As can be seen, the joint variations [governed by $\mathcal{L}_{\text{smooth}}$ and $\mathcal{L}_{\text{adv}}$ in (12)] do not change dramatically in this range, while the fooling rate [governed by $\mathcal{L}_{\text{pred}}$ in (12)] also remains acceptable.

F. Failure Cases

In our experiments, we also observed a few failure cases where the targeted model was resistant to the perturbations on some skeleton actions. These cases were more frequently observed for the advanced mode of the CIASA attack, where we constrain the attack to perturb only a few joints. Doing so results in trading off the fooling capacity with perturbation imperceptibility. In Fig. 8, we show quantitative results related to two typical failure cases. In the figure, it can be observed that prediction confidence of the model for the correct label of the original data—in Fig. 8(a)—is quite high. Consequently, targeted fooling to the least likely classes for such actions was found challenging. This was true in general regarding the failure cases in our experiments. However, our attack was mostly able to reduce the original confidence of the correct class (first row) and resulted in nontargeted fooling for the other cases (second row).

VI. Conclusion

We present the first systematic adversarial attack on skeleton-based action recognition. Unlike the existing attacks that target non-sequential tasks, e.g., image classification, semantic segmentation, and pose estimation, we attack deep sequential models from a spatiotemporal perspective. With the skeleton-based action recognition model ST-GCN [8] as the target, we demonstrate its successful fooling by mainly perturbing the joint positions. The proposed attack algorithm CIASA imposes spatiotemporal constraints on the adversarial perturbations to produce perturbed skeleton sequences with temporal smoothness, spatial integrity, and anthropomorphic plausibility. The proposed algorithm works in different modes based on the needs of the attack. With the localized mode of CIASA, we are able to perturb only a particular set of the body joints to launch the localized attack. Such attacks can be used to inject regional perturbations to prespecified parts of the body, without interfering with the dominant action patterns that are performed by the other joints. Compared to the basic mode that perturbs all the joints with global clipping, an advanced mode utilizes localized attacks with hierarchical joint variations to disguises the attack intentions with realistic motion patterns. Our experiments show that the proposed CIASA perturbations generalize well across different recognition models. Moreover, they also have the ability to transfer to RGB video modality under graphics rendering pipeline. This indicates that CIASA generated perturbations can allow attackers to mimic semantically imperceptible adversarial patterns in the real world to fool skeleton-based action recognition systems.

Acknowledgment

The GPUs used for this research are donated by the NVIDIA Corporation.

References

[1] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2752–2759.

[2] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo, "3-D human action recognition by shape analysis of motion trajectories on Riemannian manifold," *IEEE Trans. Cybern.*, vol. 45, no. 7, pp. 1340–1352, Jul. 2015.

[3] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2015, pp. 579–583.

[4] R. Vemulapalli and R. Chellappa, "Rolling rotations for recognizing human actions from 3D skeletal data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4471–4479.

[5] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3D action recognition," 2017, *arXiv:1703.03492*. [Online]. Available: http://arxiv.org/abs/1703.03492

[6] Y. Yang, C. Deng, D. Tao, S. Zhang, W. Liu, and X. Gao, "Latent max-margin multitask learning with skelets for 3-D action recognition," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 439–448, Feb. 2017.

[7] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal LSTM network with trust gates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3007–3021, Dec. 2018.

[8] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–10.

[9] C. Cao, Y. Zhang, C. Zhang, and H. Lu, "Body joint guided 3-D deep convolutional descriptors for action recognition," *IEEE Trans. Cybern.*, vol. 48, no. 3, pp. 1095–1108, Mar. 2018.

[10] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, p. 12.

[11] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1010–1019.

[12] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 28–35.

[13] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley MHAD: A comprehensive multimodal human action database," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Jan. 2013, pp. 53–60.

[14] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5344–5352.

[15] C. Szegedy *et al.*, "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*. [Online]. Available: http://arxiv.org/abs/1312.6199

[16] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*. [Online]. Available: http://arxiv.org/abs/1412.6572

[17] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroS&P)*, Mar. 2016, pp. 372–387.

[18] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 828–841, Oct. 2019.

[19] V. Mirjalili and A. Ross, "Soft biometric privacy: Retaining biometric utility of face images while perturbing gender," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 564–573.

[20] Y.-C. Lin, Z.-W. Hong, Y.-H. Liao, M.-L. Shih, M.-Y. Liu, and M. Sun, "Tactics of adversarial attack on deep reinforcement learning agents," 2017, *arXiv:1703.06748*. [Online]. Available: http://arxiv.org/abs/1703.06748

[21] N. Akhtar, M. A. A. K. Jalwana, M. Bennamoun, and A. Mian, "Label universal targeted attack," 2019, *arXiv:1905.11544*. [Online]. Available: http://arxiv.org/abs/1905.11544

[22] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*. [Online]. Available: http://arxiv.org/abs/1609.02907

[23] D. Zügner, A. Akbarnejad, and S. Günnemann, "Adversarial attacks on neural networks for graph data," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2018, pp. 2847–2856.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[25] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.

[26] W. Kay *et al.*, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*. [Online]. Available: http://arxiv.org/abs/1705.06950

[27] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018.

[28] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "VNect: Real-time 3D human pose estimation with a single RGB camera," *ACM Trans. Graph.*, vol. 36, no. 4, p. 44, 2017.

[29] Z. Huang, C. Wan, T. Probst, and L. Van Gool, "Deep learning on lie groups for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6099–6108.

[30] J. Liu, N. Akhtar, and A. Mian, "Skepxels: Spatio-temporal image representation of human skeleton joints for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 10–19.

[31] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1110–1118.

[32] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4041–4049.

[33] A. Shahroudy, T.-T. Ng, Q. Yang, and G. Wang, "Multimodal multipart learning for action recognition in depth videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2123–2129, Oct. 2016.

[34] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5323–5332.

[35] X. Gao, W. Hu, J. Tang, P. Pan, J. Liu, and Z. Guo, "Generalized graph convolutional networks for skeleton-based action recognition," 2018, *arXiv:1811.12013*. [Online]. Available: http://arxiv.org/abs/1811.12013

[36] X. Zhang, C. Xu, X. Tian, and D. Tao, "Graph edge convolutional neural networks for skeleton based action recognition," 2018, *arXiv:1805.06184*. [Online]. Available: http://arxiv.org/abs/1805.06184

[37] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Adaptive spectral graph convolutional networks for skeleton-based action recognition," 2018, *arXiv:1805.07694*. [Online]. Available: http://arxiv.org/abs/1805.07694

[38] Y.-H. Wen, L. Gao, H. Fu, F.-L. Zhang, and S. Xia, "Graph CNNS with motif and variable temporal block for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8989–8996.

[39] W. Peng, X. Hong, H. Chen, and G. Zhao, "Learning graph convolutional network for skeleton-based human action recognition by neural searching," 2019, *arXiv:1911.04131*. [Online]. Available: http://arxiv.org/abs/1911.04131

[40] B. Li, X. Li, Z. Zhang, and F. Wu, "Spatio-temporal graph routing for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8561–8568.

[41] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7912–7921.

[42] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," 2016, *arXiv:1611.02770*. [Online]. Available: http://arxiv.org/abs/1611.02770

[43] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ACM Asia Conf. Comput. Commun. Secur.*, Apr. 2017, pp. 506–519.

[44] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song, "Spatially transformed adversarial examples," 2018, *arXiv:1801.02612*. [Online]. Available: http://arxiv.org/abs/1801.02612

[45] L. Sun *et al.*, "Adversarial attack and defense on graph data: A survey," 2018, *arXiv:1812.10528*. [Online]. Available: http://arxiv.org/abs/1812.10528

[46] H. Dai *et al.*, "Adversarial attack on graph structured data," 2018, *arXiv:1806.02371*. [Online]. Available: http://arxiv.org/abs/1806.02371

[47] A. Bessi, "Two samples test for discrete power-law distributions," 2015, *arXiv:1503.00643*. [Online]. Available: http://arxiv.org/abs/1503.00643

[48] M. E. J. Newman, D. J. Watts, and S. H. Strogatz, "Random graph models of social networks," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 1, pp. 2566–2572, 2002.

[49] N. Akhtar, J. Liu, and A. Mian, "Defense against universal adversarial perturbations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3389–3398.

[50] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2794–2802.

[51] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," 2016, *arXiv:1611.01236*. [Online]. Available: http://arxiv.org/abs/1611.01236

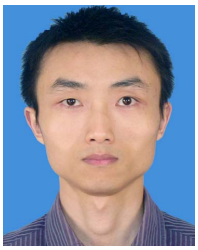
[52] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," 2016, *arXiv:1607.02533*. [Online]. Available: http://arxiv.org/abs/1607.02533

[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[54] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.

[55] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," 2018, *arXiv:1812.08008*. [Online]. Available: http://arxiv.org/abs/1812.08008

[56] J. Liu, N. Akhtar, and A. Mian, "Learning human pose models from synthesized data for robust RGB-D action recognition," 2017, *arXiv:1707.00823*. [Online]. Available: http://arxiv.org/abs/1707.00823



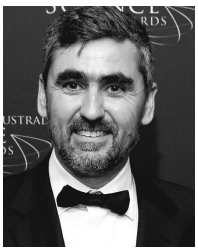
Jian Liu (Member, IEEE) received the Bachelor in Engineering degree from the Huazhong University of Science and Technology, Wuhan, China, in 2006, the master's degree from The Hong Kong University of Science and Technology, Hong Kong, in 2011, and the Ph.D. degree in computer vision from The University of Western Australia (UWA), Crawley, WA, Australia, in 2020.

His research interests include computer vision, human action recognition, deep learning, and human pose estimation.

Naveed Akhtar (Member, IEEE) received the master's degree in computer science from Hochschule Bonn-Rhein-Sieg, Sankt Augustin, Germany, in 2012, and the Ph.D. degree in computer vision from The University of Western Australia (UWA), Crawley, WA, Australia, in 2017.

He was a Research Fellow with The University of Western Australia (UWA), Crawley, WA, Australia, and Australian National University, Canberra ACT, Australia. He is currently an Assistant Professor with UWA. He has secured two research grants from the Defense Advanced Research Projects Agency (DARPA), USA. His research in computer vision and pattern recognition is regularly published in the reputed sources of his field, such as the IEEE Conference on Computer Vision and Pattern Recognition and IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. His current research interests include adversarial machine learning, explainable AI, human action recognition, 3-D point clouds, and hyperspectral image analysis.

Dr. Akhtar is serving as an Associate Editor for IEEE ACCESS and a guest editor for two other journals.



Ajmal Mian (Senior Member, IEEE) is currently a Professor of computer science with The University of Western Australia, Crawley, WA, Australia. He has secured ten Australian Research Council grants and a National Health and Medical Research Council grant. His research interests include machine learning, video analysis, and 3-D point cloud analysis.

Prof. Mian has received numerous awards including the Australasian Distinguished Doctoral Dissertation Award from the Computing Research and Education Association of Australasia, the West Australian Early Career Scientist of the Year 2012 Award, the Vice-Chancellor's Mid-Career Research Award in 2014, the IAPR Best Scientific Paper Award in 2014, the EH Thompson Award in 2015, the Aspire Professional Development Award in 2016, and the Excellence in Research Supervision Award in 2017. He received the prestigious Australian Postdoctoral and Australian Research Fellowships in 2008 and 2011. He is also an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, *Pattern Recognition* journal and IEEE TRANSACTIONS ON IMAGE PROCESSING. He has served as a Guest Editor for *Neural Computing and Applications*, *Pattern Recognition*, *Computer Vision and Image Understanding*, and *Image and Vision Computing* journals. He was the General Chair of the International Conference on Digital Image Computing Techniques and Applications (DICTA) 2019 and the Asian Conference on Computer Vision (ACCV) 2018.