Self-Supervised Learning for Multilevel Skeleton-Based Forgery Detection via Temporal-Causal Consistency of Actions

Liang Hu[*†1,2], **Dora D. Liu**[*2,3], **Qi Zhang**[†4,2], **Usman Naseem**[5], **Zhong Yuan Lai**[2]

[1]Tongji University
[2]DeepBlue Academy of Sciences
[3]BirenTech Research
[4]University of Technology Sydney
[5]University of Sydney
lianghu@tongji.edu.cn, liudongmei_0506@163.com
qi.zhang-13@student.uts.edu.au, usman.naseem@sydney.edu.au
abrikosoff@yahoo.com

Abstract

Skeleton-based human action recognition and analysis have become increasingly attainable in many areas, such as security surveillance and anomaly detection. Given the prevalence of skeleton-based applications, tampering attacks on human skeletal features have emerged very recently. In particular, checking the temporal inconsistency and/or incoherence (TII) in the skeletal sequence of human action is a principle of forgery detection. To this end, we propose an approach to self-supervised learning of the temporal causality behind human action, which can effectively check TII in skeletal sequences. Especially, we design a multilevel skeleton-based forgery detection framework to recognize the forgery on frame level, clip level, and action level in terms of learning the corresponding temporal-causal skeleton representations for each level. Specifically, a hierarchical graph convolution network architecture is designed to learn low-level skeleton representations based on physical skeleton connections and high-level action representations based on temporal-causal dependencies for specific actions. Extensive experiments consistently show state-of-the-art results on multilevel forgery detection tasks and superior performance of our framework compared to current competing methods.

Introduction

Skeleton-based human action recognition and analysis have seen wide applications in many different areas of industry and academia, in diverse fields ranging from anomaly detection (Markovitz et al. 2020) to imitation learning (Yuan and Kitani 2018; Wang et al. 2019). In recent years, an increasing number of human behavior analysis scenarios are based on *skeletal features* because it offers several advantages (Morais et al. 2019; Markovitz et al. 2020) compared to *pixel-based features*: (1) it allows the algorithm to focus on overall poses rather than irrelevant features such as illumination or background clutter; (2) the human skeleton can be represented as a compact graph that makes training and testing much faster compared to high-dimensional pixel features; (3) skeletal features are more friendly and easily transferable to robots for imitation learning.

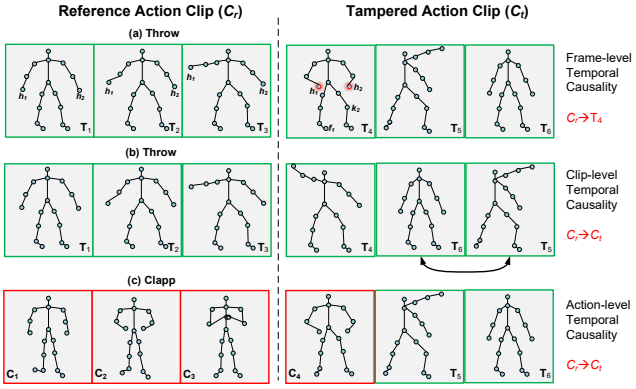

Figure 1: Frame, clip, and action-level skeleton forgeries. The joints highlighted in red are tampered from the original positions. Red and green boxes distinguish the two actions.

With the increasing skeleton-based applications, there has been emerging work on the study of direct attacks on skeletal features (Liu, Akhtar, and Mian 2022; Tanaka, Kera, and Kawamoto 2022; Diao et al. 2021). Compared to traditional pixel-based attacks, skeleton-based attacks enable *intentional misleading* of skeleton recognition systems and result in more severe consequences. For example, intentional manipulation of the skeletal action "walking" to "stop to stand" will mislead a self-driving car equipped with a skeleton-based pedestrian action recognition system. Unfortunately, there is no *skeleton-based forgery detection* (SFD) method available to our best knowledge.

Current human action recognition methods mainly consider the physical connections between skeleton joints (Yan, Xiong, and Lin 2018; Dang et al. 2021; Li et al. 2020). Johansson (Johansson 1973) showed that human actions could be recognized by only a few tracked points with specific causal continuities. Figure 1 illustrates two types of hu-

[*]These authors contributed equally.
[†]Corresponding authors

man action: *throw* and *clap*, where different actions obviously have different trajectory patterns. It is easy to find the reference action clips (a) and (c) have different temporal causalities. As a result, we can conclude T_4 is tampered because $C_r \rightarrow T_4$ violates the frame-level temporal causality. Moreover, disordering the frames (see (b)) leads to *clip-level forgery* while the insertion of frames from other actions (see (c)) leads to the *action-level forgery*. These forgeries will result in temporal inconsistency and incoherence (TII), which can be easily detected by checking if $C_r \rightarrow C_t$ violates clip- and action-level temporal causalities.

As the first attempt to tackle the skeleton-based attacks, we propose a temporal-causal SFD network (TC-SFDN) architecture to detect the forgeries at the frame, clip and action levels. We implement this architecture with hierarchical graph convolutional networks (GCNs), where low-level GCNs learn a physical skeleton embedding in terms of a *skeletal connection graph* while high-level GCNs learn a causal skeleton embedding in terms of a *temporal-causal graph* that describes logical dependencies for each action. In particular, a group of self-supervised learning (SSL) tasks are designed for multilevel SFD to avoid human labeling.

The main contributions of our work are given as follows:

- This is the first attempt of SFD, which is necessary, significant and timely to deal with the emerging challenge of the increasing skeleton-based attacks.
- We implement TC-SFDN with a hierarchical GCN architecture to learn both low-level skeleton representations based on physical body connections and high-level action representations based on the temporal-causal graph for each action instance.
- A group of SSL tasks are designed to efficiently train TC-SFDN for multilevel SFD.
- Extensive experiments are performed on two real-world datasets. Our method significantly outperforms state-of-the-art baseline models in all evaluation cases.

Related Work

Since there is no SFD method yet available, we briefly review conventional pixel-based forgery detection methods, and then we present the workaround for SFD based on current skeleton-based anomaly detection methods.

Pixel-based Forgery Detection (PFD). Forgery detection on videos has been studied for more than a decade (Hsu et al. 2008). In recent years, it has attracted more attention due to the prevalence of generative adversarial nets (Goodfellow et al. 2014), especially for video generation (Tulyakov et al. 2018). As a result, there has been increasing demand for PFD (Cozzolino et al. 2021; Javed et al. 2021). However, these PFD methods are built on low-level *pixel-based* features, which are inapplicable to emerging direct attacks on high-level *skeletal features* as addressed in this paper.

Skeleton-based Forgery Detection (SFD). Although no pre-existing work is available for SFD, there are limited studies in skeleton-based anomaly detection, which can be adapted for workarounds. Markovitz et al. (2020) use a spatial-temporal graph convolutional architecture to learn latent representations of skeletal motion trajectories; the learned representations are then subject to clustering as normal or anomalous. This work aims to distinguish anomalous action instances from normal ones instead of performing SFD within an action clip, as studied in this paper. Morais et al. (2019) propose a skeleton-based anomaly detection model, named MPED-RNN, which decomposes human motion into local and global movement elements. It measures the anomaly scores for skeleton joints in terms of the Euclidean distance between the predictive and real coordinates. Moreover, it applies a max-pooling operator over the anomaly scores of all joints as the frame anomaly score. This work is the most related to ours, and we will show how to adapt it for SFD in the experimental section.

Temporal Causality Modeling. Most current human action recognition and prediction methods mainly consider the physical skeleton connections, which are invariant to action classes with various trajectory patterns. To date, there is very limited work (Yi and Pavlovic 2012; Narayan and Ramakrishnan 2014; Tank et al. 2018) considering *temporal causality* in human action modeling. However, Johansson (1973) has shown that human actions can be recognized by a few tracked points with a specific causality. As a result, temporal causality should be considered as an essential component of detecting forgery from the skeletal trajectories of human actions.

We summarize the gaps in the above approaches to SFD.

- *Pixel-based Approach* (Cozzolino et al. 2021; Javed et al. 2021) is based on low-level pixel features that involve too much irrelevant information to the human action analysis task, which increases the burden on the model training to discriminate between signal and noise. It cannot be applied to skeleton-based applications.
- *Multivariate Time Series (MTS) Approach* (Martinez, Black, and Romero 2017; Barsoum, Kender, and Liu 2017) treats the sequence of multiple joints as a MTS prediction problem. However, it may predict skeleton joints that are inconsistent with the real body structure.
- *Physical Connection-based Approach* (Dang et al. 2021; Li et al. 2020) only models body structures but ignoring the temporal-causal difference between action classes.

Preliminaries

Problem Formalization

Let us consider SFD tasks on 3D human skeleton trajectories. A sequence $\mathcal{S}^a = \{f_1, \cdots, f_T\}$ w.r.t. human action a consists of T skeleton frames. Each $f_t \in \mathcal{S}^a$ can be described with a graph representation $f_t = \{\mathcal{V}_t, \mathcal{E}\}$ where $\mathcal{V}_t = \{\mathbf{v}_1, \cdots, \mathbf{v}_J\}$, i.e. $\mathbf{v}_j \in \mathbb{R}^3$ is the coordinate vector of joint j, and $\mathcal{E} = \{e_{ij} | (i,j) \in \mathcal{H}\}$ where $\mathcal{H}$ is the set of naturally connected body joints.

Given an action clip, $\mathcal{S}_r^a = \{f_1, \cdots, f_T\}$, as the reference, this paper aims to build an SFD model that is capable of detecting the following three-level forgeries in the subsequent clip (cf. Figure 1), $\mathcal{S}_t^a = \{f_{T+1}, \cdots, f_{T+N-1}\}$:

- **Frame-level Forgery:** Misplacing joints in a frame, i.e., a set of joints $\{\tilde{\mathbf{v}}_j\}$ being different from the original coordi-

nates $\{\mathbf{v}_j\}$ where $\tilde{\mathbf{v}}_j \neq \mathbf{v}_j$, results in an incoherent frame $\tilde{f}^t_{T+1}$ with $\mathcal{S}^a_r$.

$$\tilde{f}_{T+1}|\mathcal{S}^a_r \coloneqq \tilde{f}_{T+1} \mid f^r_T, \cdots, f^r_1 \tag{1}$$

- **Clip-level Forgery:** Changing the order of frames in $\mathcal{S}^a_t$ leads to a tampered sequence:

$$\tilde{\mathcal{S}}^a_t|\mathcal{S}^a_r \coloneqq f_{T+i}, \cdots, f_{T+N-1}, \cdots, f_{T+1} \mid \mathcal{S}^a_r \tag{2}$$

- **Action-level Forgery:** Involving skeleton frames from another sequence, i.e. $(f^b_i, f^b_j, f^b_k) \subset \mathcal{S}^b$, of action b results in an inconsistent action sequence with $\mathcal{S}^a_r$, denoted as

$$\tilde{\mathcal{S}}^a_t|\mathcal{S}^a_r \coloneqq f^a_{T+N-1}, \cdots, (f^b_i, f^b_j, f^b_k), \cdots, f^a_{T+1} \,|\mathcal{S}^a_r \tag{3}$$

Temporal-Causal Graph

As stated in the introduction, the trajectory of a human action follows specific temporal causalities (Narayan and Ramakrishnan 2014; Yi and Pavlovic 2012). To find the temporal causalities between each pair of skeleton joints, we employ transfer entropy (TE) (Schreiber 2000) because it offers some advantages: (1) TE is a model-free approach that avoids high computation cost demanded, such as Grange-causality (Narayan and Ramakrishnan 2014); (2) TE is able to detect statistical dependencies not limited to linear statistics (Schreiber 2000). In general, TE is defined as conditional mutual information with the history of the influenced variable $x_{t:t-l+1}$ in the condition.

$$T_{X\to Y} = I(y_{t+1}; y_{t:t-l+1}|x_{t:t-l+1}) \tag{4}$$

In particular, we adopt *pseudo transfer entropy* (pTE) (Silini and Masoller 2021) to compute Eq. 4 efficiently; pTE assumes that the processes X and Y follow normal distributions. Let y_{t+1} represent the state of process Y at time step $t+1$, and $y^l_t = y_{t:t-l+1}$ and $x^l_t = x_{t:t-l+1}$. We compute pTE as in (Silini and Masoller 2021):

$$T_{X\to Y} = \frac{1}{2}\log\frac{\left|\Sigma\left([\mathbf{Y}^l_t||\mathbf{X}^l_t]\right)\right| \cdot \left|\Sigma\left([\mathbf{y}_{t+1}||\mathbf{Y}^l_t]\right)\right|}{\left|\Sigma\left([\mathbf{y}_{t+1}||\mathbf{Y}^l_t||\mathbf{X}^l_t]\right)\right| \cdot \left|\Sigma\left(\mathbf{Y}^l_t\right)\right|} \tag{5}$$

where $\Sigma\left([\mathbf{X}||\mathbf{Y}]\right)$ is the covariance of the concatenation of matrices $\mathbf{X}$ and $\mathbf{Y}$, $\mathbf{y}_{t+1}$ is the vector of the future values of $\mathbf{Y}$, $\mathbf{Y}^l_t$ and $\mathbf{X}^l_t$ are matrices containing the previous l values of processes Y and X respectively.

In the context of human motion, it is straightforward to obtain the pTE, $T_{i\to j}$, between any pair of joints (i, j) according to the motion processes $\mathbf{v}_i$ and $\mathbf{v}_j$ via Eq. 5. As a result, we obtain a temporal-causal relation matrix $\mathcal{C}^a$ for each action sequence $\mathcal{S}^a$, where each element of $\mathcal{C}^a(i, j) = T_{i\to j}$ denotes joint i's temporal-causal influence on j. Note that $\mathcal{C}^a$ is asymmetrical due to the inequality between $T_{i\to j}$ and $T_{j\to i}$.

Model Specification

Figure 2 shows the architecture of our temporal-causal SFD network (TC-SFDN), where the low-level Physical Residual GCN (P-ResGCN) encodes the human motion trajectory under the constraints of physical skeleton connections while the high-level Causal Residual GCN (C-ResGCN) encodes human action according to temporal causalities specific to a particular action. We place the Temporal Reduce ConvNet (TRCN) over the output of C-ResGCN to generate an action clip representation that integrates the information over all frames. Note that the reference/target action clip encoders share the same model architecture but with different weight parameters because: (1) the input size, i.e., the number of frames in the reference action and target action clips could be different; (2) the target action clips generally contains tampered information which follows a different distribution from the reference action clips.

Physical Residual Graph ConvNet

In the P-ResGCN module, we follow the skeletal connection settings in ST-GCN (Yan, Xiong, and Lin 2018). As a result, we obtain a set of physical relation matrices, $\mathcal{A}$, with different body partition strategies, e.g., *uni-labeling*, *distance* and *spatial* (Yan, Xiong, and Lin 2018). Then, we build the GCN block with $\mathcal{A}$ as shown in Figure 2 (b):

$$\mathbf{h}_{\text{gc}} = \sum\nolimits_{\mathbf{A}_i\in\mathcal{A}} \mathbf{\Lambda}_i^{-1}\mathbf{A}_i\mathbf{h}_{\text{in}}\mathbf{W}_i \tag{6}$$

$$\mathbf{h}_{\text{out}} = \text{LeakyReLU}(\text{InstanceNorm}(\mathbf{h}_{\text{gc}})) \tag{7}$$

where $\mathbf{W}$ is the weight matrix, the input feature map $\mathbf{h}_{\text{in}} \in \mathbb{R}^{T\times V\times C}$ (T, V, C denotes the number of frames, joints, and channels) and $\mathbf{\Lambda}^{-1}$ is the diagonal matrix for random walk normalization (Kipf and Welling 2016). We use instance normalization (Ulyanov, Vedaldi, and Lempitsky 2016) to preserve the motion trajectory patterns of each action sequence, and LeakyRelu is the activation for output.

Given the reference/target clips, $\mathbf{X}_r \in \mathbb{R}^{T_r\times V\times 3}$ and $\mathbf{X}_t \in \mathbb{R}^{T_t\times V\times 3}$ (where 3 denotes the 3D coordinates) as the inputs for Reference/Target Action Clip Encoder, we stack stack N GCN blocks (Eq. 6 and 7) to form a multilayer GCN, where the input $\mathbf{X}$ could be $\mathbf{X}_r$ or $\mathbf{X}_t$:

$$\mathbf{h}_{\text{out}} = \text{GCN}^N(\mathbf{X}, \mathcal{A}) = \text{GCN}(\cdots_{\times N}\text{GCN}(\mathbf{X}, \mathcal{A})) \tag{8}$$

In this case, the representation of normal joints may integrate much error information from their tampered neighbors through multilayer GCNs. As a result, it becomes impossible to distinguish the normal and tampered joints. To avoid this issue, we use a residual connection to preserve the original information for each joint.

$$\mathbf{H} = \text{P-ResGCN}(\mathbf{X}) = \text{GCN}^N(\mathbf{X}, \mathcal{A}) + \text{T}(\mathbf{X}) \tag{9}$$

where $\text{T}(\mathbf{X})$ is a linear projection to align the dimension with $\mathbf{h}_{\text{out}}$. As a result, we obtain the *Physical Skeleton Embeddings* $\mathbf{H}_r \in \mathbb{R}^{T_r\times V\times F}$ and $\mathbf{H}_t \in \mathbb{R}^{T_t\times V\times F}$ (F is the feature size) given the inputs $\mathbf{X}_r$ or $\mathbf{X}_t$.

Causal Residual Graph ConvNet

Since human poses are constrained by the physical skeleton connections of the body, we place the C-ResGCN above the P-ResGCN (cf. Figure 2) to reveal the high-level temporal causalities of behavioral patterns behind human actions.

Following the method described in Preliminaries, we compute the temporal-causal relation matrix, $\mathcal{C}^a$, for each

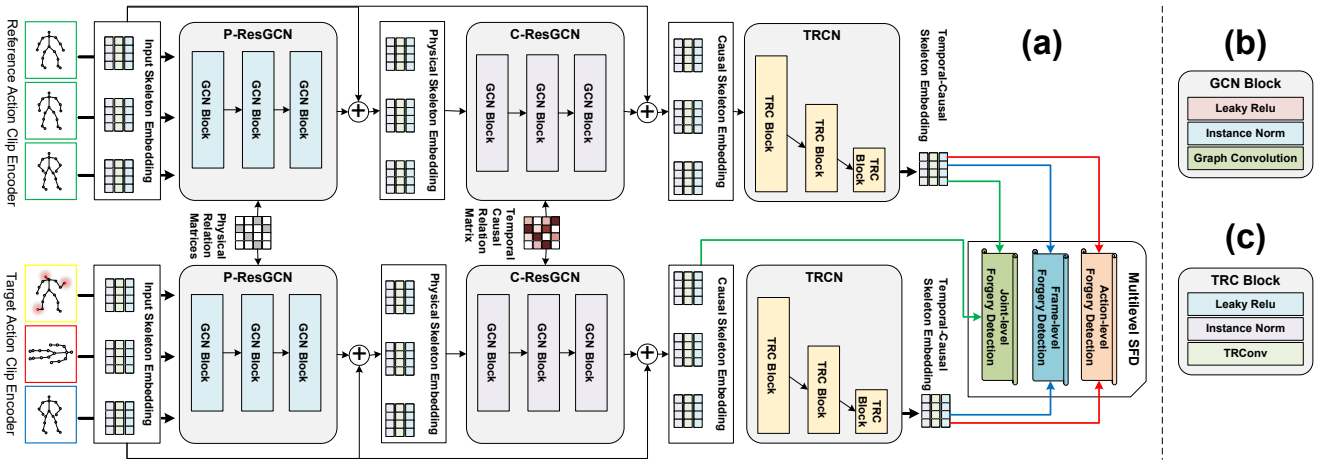

Figure 2: (a) TC-SFDN architecture, where P-ResGCNs (which stacks N-layer GCN block (b)) are designed to learn physical skeleton embeddings, C-ResGCNs are designed to learn causal skeleton embeddings, and TRCNs (stack L-layer TRC block (c)) are designed learn temporal-causal skeleton embeddings.

training action trajectory sequence, $\mathcal{S}^a$, offline. Then, we obtain the category-level temporal-causal relation matrix by the mean of all instance-level temporal-causal matrices, i.e., $\mathcal{C}^c = \text{mean}(\{\mathcal{C}^a | a \in c\})$. Given a reference action clip $\mathcal{S}_r^a$, we calculate a mixture version of the temporal-causal relation matrix as follows:

$$\bar{\mathcal{C}}_\lambda = (1-\lambda)\mathcal{C}^a + \lambda \mathcal{C}^c \tag{10}$$

where the category-level temporal-causal matrix $\mathcal{C}^c$ serves as the prior that is useful when $\mathcal{C}^a$ is not available. Specially, we set a trade-off parameter λ that takes on the value $\lambda = 1$ when the temporal-causal matrix refers only to category-level causalities. On the contrary, $\lambda = 0$ is set to ignore $\mathcal{C}^c$ when categorical information for $\mathcal{S}^r$ is not available or $\mathcal{C}^a$ does not conform to $\mathcal{C}^c$.

So far, $\bar{\mathcal{C}}_\lambda$ is a complete bidirectional temporal-causal graph over all joints. According to the principle of Occam's razor, we should only preserve the most useful dependencies because the full connections may involve noise due to the message passing mechanism of GCN. As a result, in this paper we refine $\bar{\mathcal{C}}_\lambda$ to a sparse temporal-causal matrix (Xu et al. 2020; Yi and Pavlovic 2012) preserving only 20% of the most significant temporal-causal dependency. Similar to P-ResGCN, we construct a multilayer residual GCN, namely C-ResCGN, based on the sparse temporal-causal graph $\bar{\mathcal{C}}_\lambda$:

$$\mathbf{F} = \text{C-ResGCN}(\mathbf{H}) = \text{GCN}^N(\mathbf{H}, \bar{\mathcal{C}}_\lambda) + \text{T}(\mathbf{X}) \tag{11}$$

As a result, we obtain the *Causal Skeleton Embeddings*, $\mathbf{F}_r \in \mathbb{R}^{T_r \times V \times F}$ and $\mathbf{F}_t \in \mathbb{R}^{T_t \times V \times F}$ with *Physical Skeleton Embeddings* $\mathbf{H}_r$ and $\mathbf{H}_t$ as the inputs.

Temporal Reduce ConvNet

TRCNs are designed to integrate the information over all frames to generate a temporal-causal skeleton representation. We construct temporal reduce convolution (TRC) blocks (cf. Figure 2 (c)) via 2D convolution with stride=2:

$$\mathbf{h} = \text{Conv2D}_{\text{stride}=2}(\mathbf{F}, \mathbf{W}) \tag{12}$$

$$\mathbf{h}_{\text{out}} = \text{LeakyReLU}(\text{InstanceNorm}(\mathbf{h})) \tag{13}$$

where $\mathbf{W} \in \mathbb{R}^{K \times 1}$ is the temporal kernel with size K that controls the number of neighboring frames involved. Given the input $\mathbf{F} \in \mathbb{R}^{T \times V \times F}$ with temporal length T, the convolution operation with stride 2 reduces the temporal length by half. That is, we obtain $\mathbf{h} \in \mathbb{R}^{\frac{T}{2} \times V \times F}$ when passing the first TRC block. As shown in Figure 2, we stack $M = \lceil \log_2 T \rceil$ TRC blocks to reduce the T dimension w.r.t. frames.

$$\mathbf{V} = \text{TRCN}(\mathbf{F}) = \text{TRC}^M_{\text{stride}=2}(\mathbf{F}) \tag{14}$$

Accordingly, we obtain the *Temporal-Causal Skeleton Embeddings*, $\mathbf{V}_r \in \mathbb{R}^{V \times F}$ and $\mathbf{V}_t \in \mathbb{R}^{V \times F}$ with the corresponding inputs $\mathbf{F}_r$ and $\mathbf{F}_t$.

Self-supervised Learning for Multilevel SFD

In general, it is impossible and unnecessary to label the forgeries in all types of human actions. Hence, we train our SFD model in the SSL manner to efficiently address multilevel forgery as stated in Problem Formalization. In the followings, we present how to design the SSL tasks for SFD at different levels and how to prepare the contrastive samples (Liu et al. 2021) for each SSL-based SFD task.

SFD for Frame-level Forgery. Given reference clip $\mathcal{S}_r^a$, the SFD task aims to detect if the next frame f_{T+1} is temporal incoherent with $\mathcal{S}_r^a$ (cf. Eq. 1). To learn the frame-level SFD model, we design the SSL task as follows.

SSL Task Design. One strategy to make a forged frame is randomly shifting the coordinates of each joint, but it is easily detected because of the significant violation of temporal causality. Instead, we adopt a much harder task, that is, we swap f_{T+1} to a random position in the target clip $\mathcal{S}_t^a$, and the SSL strategy is designed to find the true frame f_{T+1} with the highest temporal coherence after $\mathcal{S}_r^a$.

SFD Model and Loss. We first project the temporal-causal skeleton embedding, $\mathbf{V}_r$, and the causal skeleton embedding of frame f_{T+i} in $\mathcal{S}_t^a$, i.e., $\mathbf{F}_{T+i} = \mathbf{F}_t[T+i, :, :]$, to new spaces where $\hat{\mathbf{V}}_r, \hat{\mathbf{F}}_{T+i} \in \mathbb{R}^{V \times R}$ with the same feature

dimensionality R, where $\mathbf{W}, \mathbf{b}$ are the parameters.

$$\hat{\mathbf{V}}_r = \text{LeakyReLU}(\mathbf{V}_r\mathbf{W}_r + \mathbf{b}_r) \quad (15)$$

$$\hat{\mathbf{F}}_{T+i} = \text{LeakyReLU}(\mathbf{F}_{T+i}\mathbf{W}_t + \mathbf{b}_t)) \quad (16)$$

Then, we concatenate $\hat{\mathbf{V}}_r$ and $\hat{\mathbf{F}}_{T+i}$ and calculate the temporal coherence scores, $\mathbf{s}_f \in \mathbb{R}^V$, for all joints

$$\mathbf{s}_f = \text{MLP}([\hat{\mathbf{V}}_r, \hat{\mathbf{F}}_{T+i}]), \quad S_f = \text{Avg}(\mathbf{s}_f) \quad (17)$$

where a 3-layer MLP (multilayer perceptron) is used to map the input features to scores, and we take the average score over all joints to measure the temporal coherence between frame f_{T+i} and reference clip $\mathcal{S}_r^a$. The same way, we obtain the score S_f for each frame f in the swapped clip $\bar{\mathcal{S}}_t^a$.

As a result, we minimize the categorical cross entropy (CCE) loss to maximize the probability $p_{\hat{f}}$ of frame $\hat{f}$ which corresponds to the first frame f_{T+1} in the original $\mathcal{S}_t^a$.

$$\{p_f\} = \text{SoftMax}(\{s_f | f \in \bar{\mathcal{S}}_t^a\}) \quad (18)$$

$$L_F = \text{CCE}(p_{\hat{f}} | \{p_f\}) \quad (19)$$

SFD for Clip-level Forgery. This SFD task aims to check if the order of frames of target sequence $\mathcal{S}_t^a$ are tampered (cf. Eq. 2), which is implemented by measuring the TII score between $\mathcal{S}_r^a$ and $\mathcal{S}_t^a$.

SSL Task Design. We randomly disorder N consecutive frames from a target clip $\mathcal{S}_t^a$. This tampered clip $\bar{\mathcal{S}}_t^a$ is labeled as a negative sample, while the original $\mathcal{S}_t^a$ is labeled as the positive one for training the SFD model.

SFD Model and Loss. Since the temporal-causal skeleton embeddings $\mathbf{V}_r$ and $\mathbf{V}_t$ (cf. Eq. 14) serve as the temporal-causal representations w.r.t. $\mathcal{S}_r^a$ and $\mathcal{S}_t^a$ respectively, we measure their TII score as follows,

$$\hat{\mathbf{V}}_t = \text{LeakyReLU}(\mathbf{V}_t\mathbf{W}_t + \mathbf{b}_t) \quad (20)$$

$$\bar{\mathbf{V}}_r = \hat{\mathbf{V}}_r / ||\hat{\mathbf{V}}_r||, \quad \bar{\mathbf{V}}_t = \hat{\mathbf{V}}_t / ||\hat{\mathbf{V}}_t|| \quad (21)$$

$$\mathbf{S}_c = \text{Tanh}([\bar{\mathbf{V}}_r, \bar{\mathbf{V}}_t]\mathbf{W}_c + \mathbf{b}_c) \quad (22)$$

$$\mathbf{s}_c = \text{AvgPooling}(\mathbf{S}_c) \quad (23)$$

where $\hat{\mathbf{V}}_r$ is obtained from Eq. 15, and $\mathbf{s}_c \in \mathbb{R}^{F_c}$ denotes the feature vector to represent TII between $\mathbf{V}_r$ and $\mathbf{V}_t$. The loss of SFD for clip-level forgery is jointly modeled with SFD for action-level forgery, as presented below.

SFD for Action-level Forgery. This SFD task aims to find if the target clip $\mathcal{S}_t^a$ contains a sequence of frames $\{f_i, \cdots, f_{i+K}\} \subset \mathcal{S}^b$ of another action, which can be checked by modeling a TII score analogous to the above.

SSL Task Design. We randomly select N frames from the target clip $\mathcal{S}_t^a$ and replace them with the frames sampled from other action clip $\mathcal{S}^b$. This tampered clip $\bar{\mathcal{S}}_t^a$ is labeled as a negative sample, while the original $\mathcal{S}_t^a$ is labeled as the positive one for training the SFD model.

SFD Model and Loss. Since both this task and the above task are identical to check the TII between reference clip $\mathcal{S}_r^a$ and target clip $\mathcal{S}_t^a$, we use the same SFD model (cf. Eq. 20-23) to measure TII scores for both tasks. Note that the action-level forgery tends to have a higher severity of TII than that caused by clip-level forgery. As a result, we differentiate them in terms of TII severity with the ordinal labels:

$$\mathcal{L} = \{\text{Normal} : 0,\ \text{Clip Forgery} : 1,\ \text{Action Forgery} : 2\}$$

Accordingly, this leads to an ordinal classification problem which can be learned by minimizing the following weighted Kappa loss (de la Torre, Puig, and Valls 2018).

$$[P_0, P_1, P_2] = \text{SoftMax}(\mathbf{s}_c^\top \mathbf{W}_p + \mathbf{b}_p) \quad (24)$$

$$L_T = \text{WeightedKappa}([P_0, P_1, P_2]) \quad (25)$$

where $\mathbf{W}_p \in \mathbb{R}^{F_c \times 3}$ maps the feature vector $\mathbf{s}_c$ to the probabilities of the above three ordinal labels, i.e. P_0, P_1, P_2.

Implementation and Training

We implement our model with Tensorflow 2.0, and use Adam (Kingma and Ba 2017) as the optimizer to minimize the following loss combining L_F (Eq. 19) and L_T (Eq. 25). We train our model on a server with GTX 1080Ti and 128G memory.

$$L = L_F + L_T \quad (26)$$

Experiments

Data Preparation

Datasets. The following two real-world datasets with rich action classes are used for the empirical study:

NTU-RGB+D (Liu et al. 2020): This large-scale dataset for RGB+D human action recognition contains 60 action classes, including daily, mutual, and health-related activities. The 3D skeleton data in this dataset consists of 25 major body joints.

PKU-MMD (Liu et al. 2017): This large-scale dataset for 3D human action understanding covers a wide range of complex human activities. PKU-MMD contains 1076 long video sequences in 51 action categories. It contains almost 20,000 action instances.

Training/Testing Sets. We sample a clip with 70 consecutive frames from each human action trajectory in NTU-RGB+D and PKU-MMD, where 80% sampled instances are used for training and the remaining 20% for testing. Then, we randomly use 50% of them to construct the forgery instances. More specifically, we split a selected clip into 50 and 20 frames, where the first 50 frames are taken as the reference action clip, and the remaining 20 frames are used as the target action clip. The forgery instances are constructed as presented in *SSL Task Design* for each level:

- *Frame-level forgery instances.* We swap the first frame in the target clip to a random position.
- *Clip-level forgery instances.* We randomly select 10 frames from the target clip and disorder them.
- *Action-level forgery instances.* We select different proportions of frames ($25\%, 50\%, 75\%$, i.e., $5, 10, 15$, frames) from the target clip and replace them with the frames sampled from other action categories.

Comparison Settings

Metrics. A collection of metrics are applied for comprehensive evaluation, including Accuracy (ACC), Area under ROC Curve (AUC) and Mean Reciprocal Rate (MRR).

Method	eat meal		pick up		stand up		clapping		drink water	back pain	writing	put on shoe
	ACC	MRR	ACC	MRR	ACC	MRR	ACC	MRR	ACC	ACC	ACC	ACC
ResRNN	0.1357	0.2541	0.1185	0.2182	0.2422	0.3599	0.0794	0.2107	0.1992	0.0898	0.0996	0.1406
MSR-GCN	0.1362	0.2800	0.1964	0.3028	0.2019	0.3261	0.0870	0.2363	0.1349	0.1438	0.0593	0.1230
HP-GAN	0.1985	0.3146	0.1822	0.3047	0.1744	0.2964	0.1761	0.2998	0.1423	0.1250	0.1414	0.1445
MPED-RNN	0.1537	0.2814	0.1621	0.2657	0.3920	0.4696	0.0971	0.2316	0.1531	0.0850	0.1014	0.1298
TC-SFDN**	**0.4538**	**0.3405**	**0.6948**	**0.4274**	**0.7778**	**0.4911**	**0.3747**	**0.3016**	**0.3634**	**0.4626**	**0.3295**	**0.5356**

Table 1: Comparison of ACC and MRR for next frame prediction on the first four action categories and only ACC results are reported for the last four categories due to the space limitation. ** indicates significance with Wilcoxon signed-rank at α=0.01.

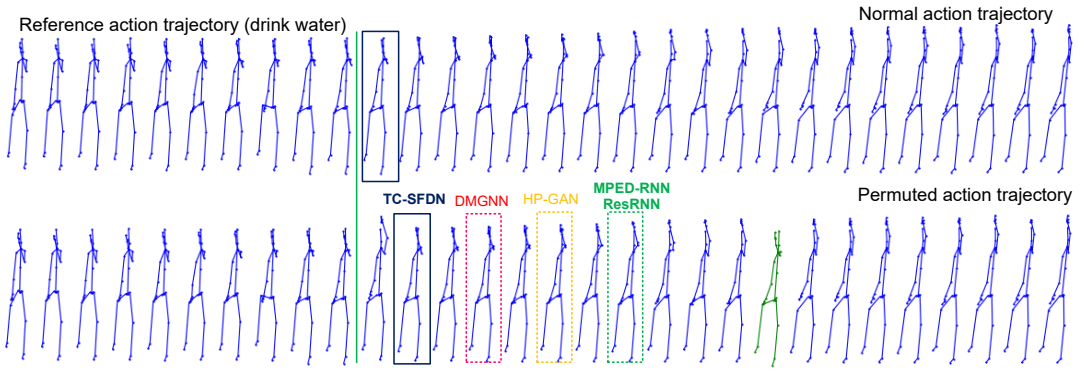

Figure 3: A demonstration of next frame prediction by baseline methods and TC-SFDN (with differently colored boxes)

Experimental Results

Baselines. As stated in the introduction, this work is the first attempt at multilevel SFD tasks. Therefore, no direct baselines are available. According to the related work section, MPED-RNN (Morais et al. 2019) is the most similar to our framework in that it measures the anomaly scores for joints and frames by the discrepancy between predictions and actual skeleton instances. More specifically, given joint v_j in future frame f_{T+i}, its anomaly score S_{v_j} is proportional to the squared Euclidean distance between the predictive coordinate $\tilde{\mathbf{v}}_j$ and the observed one $\mathbf{v}_j$:

$$S_{v_j} \coloneqq ||\tilde{\mathbf{v}} - \mathbf{v}_j||_2^2 \tag{27}$$

According to (Morais et al. 2019), a max-pooling operator is applied to aggregate the anomaly scores of all joints to obtain the frame-level anomaly score S_f,

$$S_f \coloneqq \max\{S_{v_j} | v_j \in f_{T+i}\} \tag{28}$$

which can be viewed as the counterpart of the frame-level score (cf. Eq. 17) applied to our model. Similarly, we easily obtain the mean anomaly score over all frames in target clip $\mathcal{S}_t^a$ to detect the clip/action-level forgeries.

$$S_c \coloneqq \text{mean}\{S_f | f \in \mathcal{S}^t\} \tag{29}$$

Most current skeleton-based methods focus on *human action recognition* whereas the available *human motion prediction* methods that can be adapted for SFD are much fewer. In particular, we adapt three representative of them, *ResRNN* (Martinez, Black, and Romero 2017) (MTS-based), *HP-GAN* (Barsoum, Kender, and Liu 2017) (GAN-based), and *MSR-GCN* (Dang et al. 2021) (GCN-based), to our problem by computing the anomaly scores S_f (cf. Eq. 28) and S_c (cf. Eq. 29) based on their predictive joints.

Frame-level Forgery Detection. Table 1 shows the ACC and MRR for next frame prediction on NTU-RGB+D w.r.t. eight action categories. Note that a random selection from the target clips should yield an ACC of 0.05 (1/20). With this in mind, our model performs quite well in picking the correct frame with the highest temporal coherence (cf. Eq. 17). In comparison, all the baselines significantly underperform our model. Figure 3 visually reveals the cause behind these results. From the original action trajectory, we find that near future frames in the target clips are very similar, and hence it is hard to distinguish their minute differences according to the coordinates. As a result, all the baselines identify similar frames to the correct frame (with the blue box). Our model successfully detects the correct next frame thanks to the temporal-causal dependencies between frames instead of the joint-wise predictive discrepancy score (cf. Eq. 28).

Clip-level Forgery Detection. Table 2 compares the AUC of disordered frame detection, where the AUC results show the probability of a method successfully identifying those disordered clips from the normal ones. According to Table 2, our model achieves significantly better detection performance compared to the baselines measuring the forgery in terms of predictive discrepancy (cf. Eq. 29). The main reason leading to the underwhelming performance of these baselines is that the prediction accuracy on future frames generally decreases with the increase of future steps (i.e.,

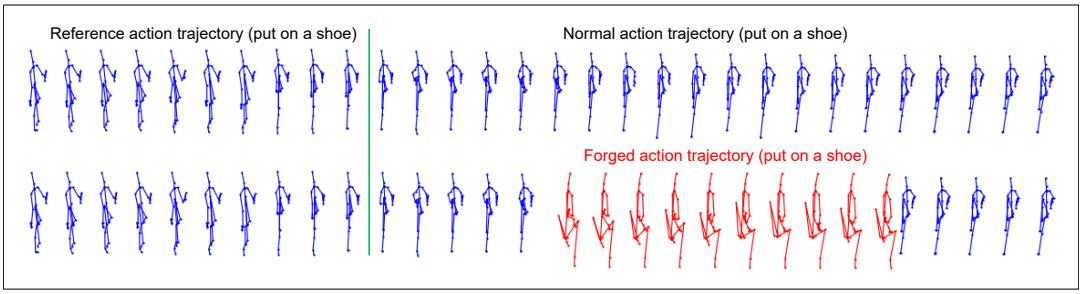

Figure 4: A demonstration of forgery on an action clip about *put on a shoe*. The AUCs of SFD on the action category *put on a shoe* are: ResRNN (0.5920), MSR-GCN (0.6515), HP-GAN (0.5931), MPED-RNN (0.5337), TC-SFDN (0.7281).

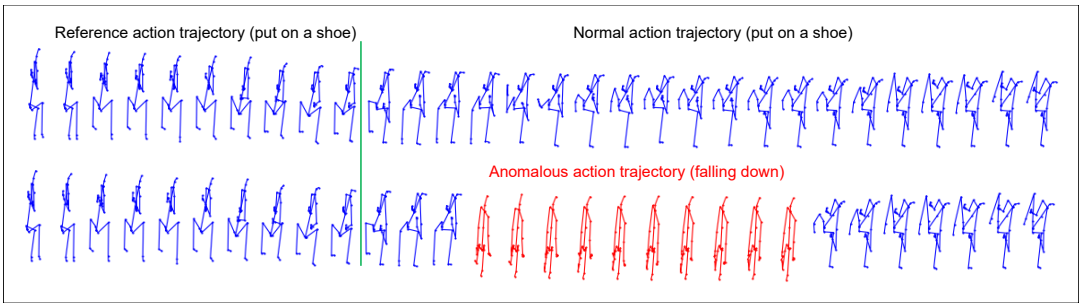

Figure 5: An example of anomalous behavior detection for *falling down* occurring in an action sample of *put on a shoe*. The AUCs of anomaly detection on the *falling down* test cases are: ResRNN (0.7034), MSR-GCN (0.7577), HP-GAN (0.7444), MPED-RNN (0.6359), TC-SFDN (0.7637).

with i increase in Eq. 28), and the neighbor frames in an action clip are very close. Therefore, all those predictive discrepancy-based methods fail in this task. In comparison, TC-SFDN detects the disorder of frames according to the TII score by checking if the temporal causality is violated, which avoids the shortcomings of predictive discrepancy measurements.

Method	drink water	back pain	writing	eat meal
ResRNN	0.4573	0.5013	0.5087	0.4413
MSR-GCN	0.4828	0.4168	0.4023	0.4529
HP-GAN	0.4785	0.5404	0.5009	0.5085
MPED-RNN	0.5122	0.5269	0.5015	0.4910
TC-SFDN**	**0.8527**	**0.8977**	**0.7539**	**0.8288**

Table 2: Comparison of the AUCs for clip-level forgery detection

Action-level Forgery Detection. Table 3 shows the AUC of the detection on action-level tampering, where the different percentage of frames in a target clip is replaced with the frames sampled from other actions. As expected, the performance decreases with a larger percentage of frames being replaced. TC-SFDN significantly outperforms the other models for the action forgery detection tasks, which again proves that it efficiently identifies temporal inconsistency by the temporal causalities. Figure 4 shows an example of forgery detection w.r.t. a specific action category, namely *put on a shoe*. We find TC-SFDN successfully detects the occurrence of action-level forgery with the highest AUC 0.7281.

Methods	NTU-RGB+D			PKU-MMD	
	25%	50%	75%	25%	50%
ResRNN	0.6078	0.7003	0.7380	0.5125	0.6496
MSR-GCN	0.6821	0.6721	0.7019	0.4453	0.6441
HP-GAN	0.6705	0.6887	0.7034	0.4828	0.6710
MPED-RNN	0.5580	0.5898	0.5808	0.5785	0.5835
TC-SFDN**	**0.7672**	**0.8199**	**0.8507**	**0.7864**	**0.8217**

Table 3: Comparison of the AUCs for action-level forgery detection

Side Effect for Anomalous Behavior Detection. The capability of SFD on action-level forgery is possibly applied to some scenarios of anomalous behavior detection. More specifically, it can be utilized to detect anomalous behavior that is suddenly inconsistent with the previous action trajectory. Figure 5 demonstrates an example of "*falling down*" (anomalous behavior) occurring when a user is putting on a shoe (normal behavior). TC-SFDN achieves the highest AUC (0.7637) in this case due to the temporal-causality modeling, which is largely beneficial for checking for temporal inconsistency.

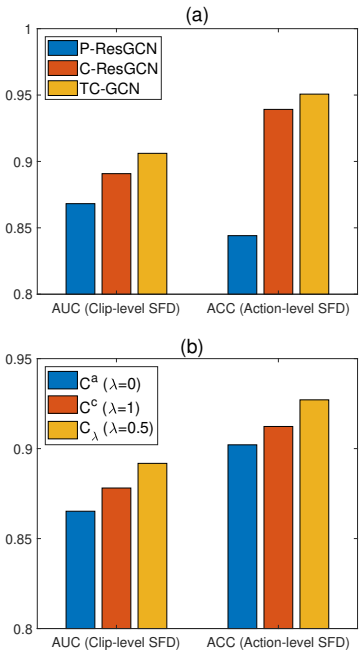

Figure 6: (a) Effects of physical connections and temporal-causal dependencies; (b) Effects of instance-level and causal-level temporal causality.

Ablation Studies

In this section, we report on ablation studies highlighting individual aspects of our model. We ablate with regards to both *network architectures* and *causality settings*. For both these cases, we compute *AUC* for frame-level forgery detection and *ACC* for action-level forgery detection.

Physical Graph vs Temporal-Causal Graph. The two main components of TC-SFDN are the P-ResGCN and C-ResGCN. For this ablation study, we retain, in turn, P-ResGCN or C-ResGCN and compare their SFD performance with that of the full model. Fig 6 (a) compares their *AUC* for frame-level forgery detection and *ACC* for action-level forgery detection. We find C-ResGCN outperforms P-ResGCN because P-ResGCN only models physical connections, which fails to capture the difference in temporal-causal motion patterns behind different actions categories. The full model slightly outperforms C-ResGCN, which proves that the regularization by physical body connections benefits the high-level temporal-causal modeling.

Category-level vs Instance-level Temporal Causality. Figure 7 illustrates three instances of temporal-causal relation matrices w.r.t. action category *put on a shoe*, where each column of the temporal-causal matrix visualizes the strengths of temporal-causal dependency from each joint to a target joint. We easily find the difference between each instance-level temporal-causal matrix, which enables to capture specialized temporal causality in each action clip.

By setting different configuration of λ (cf. Eq. 10), we obtain different temporal-causal relationships. In this study,

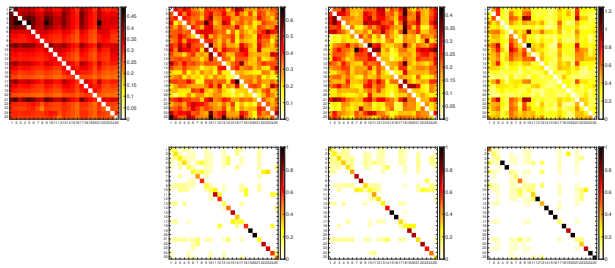

Figure 7: Visualization of temporal-causal relation matrices: (1) category-level temporal-causal matrix (row 1 and column 1); (2) instance-level temporal-causal matrices w.r.t. three action clips (row 1 and column 2-4); (3) sparse mixture temporal-causal matrix with $\lambda = 0.5$ and preserving top 20% significant values (row 2 and column 2-4).

we compare the setting of λ=0 for $\mathcal{C}^a$ only, λ=1 for $\mathcal{C}^c$ only, and λ=0.5 for $\bar{\mathcal{C}}_\lambda$ as a mixture. From the comparison in Figure 6 (b), we find that the performance using the category-level temporal-causal matrix $\mathcal{C}^c$ is better than that of $\mathcal{C}^a$, which reflects the fact that motion trajectories of the same action are closely related to each other in terms of causality (Johansson 1973; Narayan and Ramakrishnan 2014). Obviously, applying $\bar{\mathcal{C}}_\lambda$ achieves the best performance. This is because the mixed temporal-causal matrix $\bar{\mathcal{C}}_\lambda$ incorporates temporal causal dependencies from $\mathcal{C}^a$, which partially enables to model those long-tail action instances with biased causalities from the mean temporal causality, i.e. $\mathcal{C}^c$.

Conclusions

This paper initiates the first attempt to address the emerging challenge of skeleton-based forgery attacks. To this end, we propose a multilevel SFD architecture, namely TC-SFDN, where temporal causalities are introduced to capture action-specific patterns as well as the physical skeleton connections for traditional human motion modeling. Extensive empirical studies show that TC-SFDN significantly outperforms baselines in the frame, clip, and action-level SFD tasks. Future extensions of our work include adversarial attack detection and skeletal motion generation for robot control.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Granted No. 62276190, 62076184, 61976158, 61976160, 62076182) and in part by the Shanghai Innovation Action Project of Science and Technology (20511100700) and Shanghai Natural Science Foundation (22ZR1466700).

References

Barsoum, E.; Kender, J.; and Liu, Z. 2017. Hpgan: Probabilistic 3d human motion prediction via gan. 2018 IEEE. In *CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1499–149909.

Cozzolino, D.; Rossler, A.; Thies, J.; Nießner, M.; and Verdoliva, L. 2021. Id-reveal: Identity-aware deepfake video

detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15108–15117.

Dang, L.; Nie, Y.; Long, C.; Zhang, Q.; and Li, G. 2021. MSR-GCN: Multi-Scale Residual Graph Convolution Networks for Human Motion Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11467–11476.

de la Torre, J.; Puig, D.; and Valls, A. 2018. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognition Letters*, 105: 144–154. Machine Learning and Applications in Artificial Intelligence.

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

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Hsu, C.-C.; Hung, T.-Y.; Lin, C.-W.; and Hsu, C.-T. 2008. Video forgery detection using correlation of noise residue. In *2008 IEEE 10th workshop on multimedia signal processing*, 170–174. IEEE.

Javed, A. R.; Jalil, Z.; Zehra, W.; Gadekallu, T. R.; Suh, D. Y.; and Piran, M. J. 2021. A comprehensive survey on digital video forensics: Taxonomy, challenges, and future directions. *Engineering Applications of Artificial Intelligence*, 106: 104456.

Johansson, G. 1973. Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, 14(2): 201–211.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Li, M.; Chen, S.; Zhao, Y.; Zhang, Y.; Wang, Y.; and Tian, Q. 2020. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 214–223.

Liu, C.; Hu, Y.; Li, Y.; Song, S.; and Liu, J. 2017. PKU-MMD: A Large Scale Benchmark for Skeleton-Based Human Action Understanding. In *Proceedings of the Workshop on Visual Analysis in Smart and Connected Communities*, VSCC '17, 1–8. New York, NY, USA: Association for Computing Machinery. ISBN 9781450355063.

Liu, J.; Akhtar, N.; and Mian, A. 2022. Adversarial Attack on Skeleton-Based Human Action Recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 33(4): 1609–1622.

Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.-Y.; and Kot, A. C. 2020. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10): 2684–2701.

Liu, X.; Zhang, F.; Hou, Z.; Mian, L.; Wang, Z.; Zhang, J.; and Tang, J. 2021. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*.

Markovitz, A.; Sharir, G.; Friedman, I.; Zelnik-Manor, L.; and Avidan, S. 2020. Graph embedded pose clustering for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10539–10547.

Martinez, J.; Black, M. J.; and Romero, J. 2017. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2891–2900.

Morais, R.; Le, V.; Tran, T.; Saha, B.; Mansour, M.; and Venkatesh, S. 2019. Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11996–12004.

Narayan, S.; and Ramakrishnan, K. R. 2014. A cause and effect analysis of motion trajectories for modeling actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2633–2640.

Schreiber, T. 2000. Measuring information transfer. *Physical review letters*, 85(2): 461–464.

Silini, R.; and Masoller, C. 2021. Fast and effective pseudo transfer entropy for bivariate data-driven causal inference. *Scientific reports*, 11: 8423/1–8423/13.

Tanaka, N.; Kera, H.; and Kawamoto, K. 2022. Adversarial Bone Length Attack on Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2335–2343.

Tank, A.; Covert, I.; Foti, N.; Shojaie, A.; and Fox, E. 2018. Neural Granger Causality. *arXiv preprint arXiv:1802.05842*.

Tulyakov, S.; Liu, M.-Y.; Yang, X.; and Kautz, J. 2018. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1526–1535.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. S. 2016. Instance Normalization: The Missing Ingredient for Fast Stylization. *CoRR*, abs/1607.08022.

Wang, B.; Adeli, E.; Chiu, H.; Huang, D.; and Niebles, J. C. 2019. Imitation Learning for Human Pose Prediction. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 7123–7132. IEEE.

Xu, H.; Huang, Y.; Duan, Z.; Wang, X.; Feng, J.; and Song, P. 2020. Multivariate Time Series Forecasting with Transfer Entropy Graph. arXiv:2005.01185.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*.

Yi, S.; and Pavlovic, V. 2012. Sparse granger causality graphs for human action classification. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 3374–3377. IEEE.

Yuan, Y.; and Kitani, K. M. 2018. 3D Ego-Pose Estimation via Imitation Learning. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, volume 11220 of *Lecture Notes in Computer Science*, 763–778. Springer.