Fake It Till You Recognize It: Quality Assessment for Human Action Generative Models

Bruno Degardin, Vasco Lopes, and Hugo Proença, *Senior Member, IEEE*

***Abstract*—Skeleton-based generative modelling is an important research topic to mitigate the heavy annotation process. In this work, we explore the impact of synthetic data on skeleton-based action recognition alongside its evaluation methods for more precise quality extraction. We propose a novel iterative weakly-supervised learning generative strategy for synthesising high-quality human actions. We combine conditional generative models with Bayesian classifiers to select the highest-quality samples. As an essential factor, we designed a discriminator network that, together with a Bayesian classifier relies on the most realistic instances to augment the amount of data available for the next iteration without requiring standard cumbersome annotation processes. Additionally, as a key contribution to assessing the quality of samples, we propose a novel measure based on human kinematics instead of employing commonly used evaluation methods, which are heavily based on images. The rationale is to capture the intrinsic characteristics of human skeleton dynamics, thereby complementing model comparison and alleviating the need to manually select the best samples. Experiments were carried out over four benchmarks of two well-known datasets (NTU RGB+D and NTU-120 RGB+D), where both our framework and model assessment can notably enhance skeleton-based action recognition and generation models by synthesising high-quality and realistic human actions.**

***Index Terms*—Weakly-supervised learning, self-supervised learning, graph convolutional networks, generative adversarial networks, skeleton-based action recognition.**

I. Introduction

HUMAN behaviour analysis through skeleton-based data has been a crucial area of research for many years [10], [40], [44], [56], [60]. The emergence of deep learning techniques sparked a growing interest in the field, particularly due to the remarkable robustness of skeleton data in dealing with dynamic circumstances, variations, and cluttered backgrounds. Nevertheless, the performance of data-driven approaches is heavily dependent on the amount of learning data available, which is where synthetic data generation comes in as a solution to address the problem of limited data.

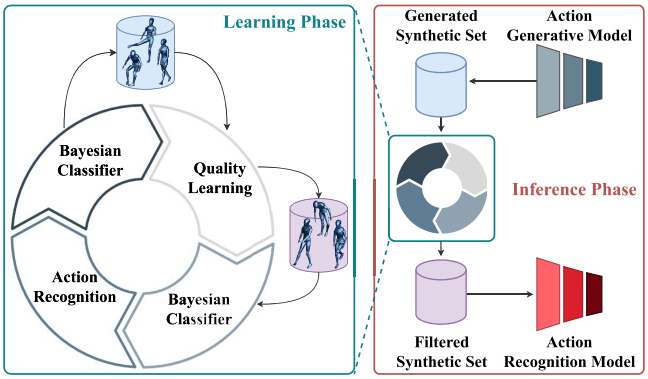


Fig. 1. **Weakly-supervised Strategy**. The proposed method builds upon a weakly supervised discriminator network that distinguishes between well and poorly-generated skeleton sequences. The results of this network are then fed into a Bayesian classification module, which selects the most confidently classified instances from an unsupervised set. These elements are consistently used as input for an action recognition model in an iterative and self-supervised way.

One of the biggest impediments to future research is the lack of quantitative evaluation methods to assess the quality of trained models accurately. Current methods to generate synthetic human actions have several limitations considering the global movement [5], [53], [55], [58], restricted ability to control the synthesised actions [21], [50], [53], [58] and the generation of high-quality samples. The latter still requires considerable human confirmation to generate the best samples, mainly due to the conventional evaluation metrics being originally proposed for image-based generative models [22], [39], [47]. Commonly used evaluation methods for human action synthesis, such as the Fréchet Inception Distance (FID) [22], where the Inception network [48] is often replaced with a skeleton-based classifier, correlate well with the perceived quality of samples and are quite sensitive to mode dropping [34], [38]. However, such metrics cannot distinguish between different failure cases (human action sequences poorly performed) since they only yield one-dimensional scores.

Manuscript received 26 July 2023; revised 29 December 2023; accepted 23 February 2024. Date of publication 12 March 2024; date of current version 3 April 2024. This work was supported in part by FCT/MEC through National Funds; in part by the FEDER-PT2020 Partnership Agreement under Project CENTRO-01-0247-FEDER-113023 (DeepNeuronic); in part by FCT/MCTES through the National Funds and Co-Funded by EU Funds under Project UIDB/50008/2020; in part by the 'Fundação para a Ciência e Tecnologia (FCT)' under Research Grant UI/BD/150765/2020 and Grant 2020.04588.BD. This article was recommended for publication by Associate Editor Z. Zhang upon evaluation of the reviewers' comments. *(Corresponding author: Bruno Degardin.)*

Bruno Degardin is with the Instituto de Telecomunicações, 3810-193 Aveiro, Portugal, also with DeepNeuronic, 6200-284 Covilhã, Portugal, and also with the Computer Science Department, University of Beira Interior, 6201-001 Covilhã, Portugal (e-mail: bruno.degardin@ubi.pt).

Vasco Lopes is with DeepNeuronic, 6200-284 Covilhã, Portugal, and also with the Computer Science Department, University of Beira Interior, 6201-001 Covilhã, Portugal (e-mail: vasco.lopes@ubi.pt).

Hugo Proença is with the Instituto de Telecomunicações, 3810-193 Aveiro, Portugal, and also with the Computer Science Department, University of Beira Interior, 6201-001 Covilhã, Portugal (e-mail: hugomcp@di.ubi.pt).

Digital Object Identifier 10.1109/TBIOM.2024.3375453

The lack of optimal quality filtration has a consequential impact on one of the key applications of human action synthesis: improving human action recognition models through synthetic data. Data augmentation methods can be broadly classified into handcrafted and reconstructive approaches. Handcrafted methods apply 3D transformations [26], [51], such as rotation, scaling, shear, horizontal flipping [57], and the addition of Gaussian noise [29]. On the other hand, reconstructive approaches [49], [54] rely on techniques such as LSTM and autoregressive methods to generate new sequences by learning temporal dynamics. However, these techniques may not capture the action's overall spatial and temporal view since they are based on image augmentation techniques. In contrast, our work focuses on exploiting generative adversarial and graph convolutional networks to overcome these limitations and capture both the data's structural and temporal aspects.

This work extends the use of Graph Convolutional-based GANs to improve skeleton-based human action recognition. We leverage the advantages of both GANs and Graph Convolutional Networks (GCNs) to overcome the limitations of current methods. Furthermore, in order to enhance the overall performance, we derive a novel notion of assessing synthetic data, where we filter the best samples by leveraging human kinematics principles related to the velocity and acceleration of human skeleton actions, explicitly focusing on metrics such as JERK and JITTER. Additionally, we further adopt an iterative learning strategy based on a weakly-supervised paradigm, where we generate conditioned human action sequences by considering the quality of the human structure and action motion. Our approach facilitates the generation of high-quality human action sequences, resulting in improved action recognition performance. This is achieved without the need for manual labeling. Notably, our method enables the automatic selection of the best-quality samples, independently from the real data.

In conclusion, this work provides several advancements: 1) the use of a scalable weakly-supervised learning discriminator framework for conditioning and filtering synthetic data to enhance human action recognition models, 2) the ability to extend the architecture to a conditional model, capable of generating a wide range of actions (up to 120 different classes), 3) a novel methodology that evaluates human generative models based on the intrinsic characteristics of the data, and 4) an extensive evaluation of the proposed work on four benchmarks from prominent datasets: NTU RGB+D [42] and NTU-120 RGB+D [30], demonstrating significant improvements over the state-of-the-art across action synthesis and action recognition.

II. Related Work

A. Skeleton-Based Behaviour Analysis

One promising cue for human behavior analysis is body pose estimation. By representing human dynamics in the form of body poses, the information extracted is semantically rich and highly descriptive, which allows reducing the impact of appearance noise that is commonly present in RGB and depth data. Thus, allowing the learning process to focus solely on human behavior.

In the past decade, skeleton-based behavior analysis has evolved significantly. This progression has ranged from using pseudo-images with CNNs [25], [33], [46] and sequence coordinate vectors with RNNs [15], [31], [32], [45], [59] to the recent adoption of GCNs [7], [8], [9], [11], [12], [44], [56], [60]. The employment of GCNs models skeleton data as spatiotemporal graphs, providing more effective capture of underlying structural information.

B. Human Action Synthesis

Skeleton-based human action synthesis can be categorized into two main approaches: autoregressive and generative methods. Autoregressive methods [16], [61] utilize Recurrent Neural Networks (RNNs) to model actions by considering skeleton data as a sequential vector of multiple frames. In contrast, generative methods leverage Generative Adversarial Networks (GANs) [18] to produce a full-body skeleton sequence from the latent space. Generative models address the problems of suboptimal extraction of structural body information and limited scalability in terms of bidirectional temporal dependency by generating synthetic data from the latent space. However, some generative methods still rely on autoregressive techniques [28], [53], [58] and Gaussian processes [55] to solve long-term relationships in the latent space and use manually structured vector sequences to model skeleton data, which can limit their scalability for action conditioning. To overcome these limitations, recent approaches [13], [20] leverage the properties of GANs and GCNs to enable effective and scalable action conditioning.

C. Evaluation of Skeleton Generative Models

Both classic and recent approaches in the field have traditionally relied on assessing the likelihood or statistical divergence of the entire set of synthetic samples when compared to the real dataset. Most evaluation metrics used to evaluate the quality of skeleton-based data are re-appropriations of metrics originally proposed for image data. Examples include the Inception Score (IS) [39] and Fréchet Inception Distance (FID) [22]. The IS provides a quantitative evaluation of the quality of generated samples within the data context. It measures the balance between the conditional label distribution $p(y|x)$ and the label distribution over the entire dataset $p(y)$. The conditional label distribution $p(y|x)$ indicates the meaningful information contained in the samples and should have low entropy, while the label distribution $p(y)$ should exhibit high entropy. As the score relies on a classifier, it requires a labelled dataset, which has been observed to be insufficient in providing reliable guidance for model comparison [2]. In the opposite spectrum, the FID is proposed as an alternative approach that does not require labelled data. The samples are embedded in a feature space, typically using a spatiotemporal graph convolution network, and then fitted with a continuous multivariate Gaussian distribution. Finally, the distance is computed as $FID(x, g) = \|\mu_x - \mu_g\|_2^2 + Tr(\Sigma_x + \Sigma_g - 2(\Sigma_x\Sigma_g)^{\frac{1}{2}})$, where μ and Σ are the mean and covariance

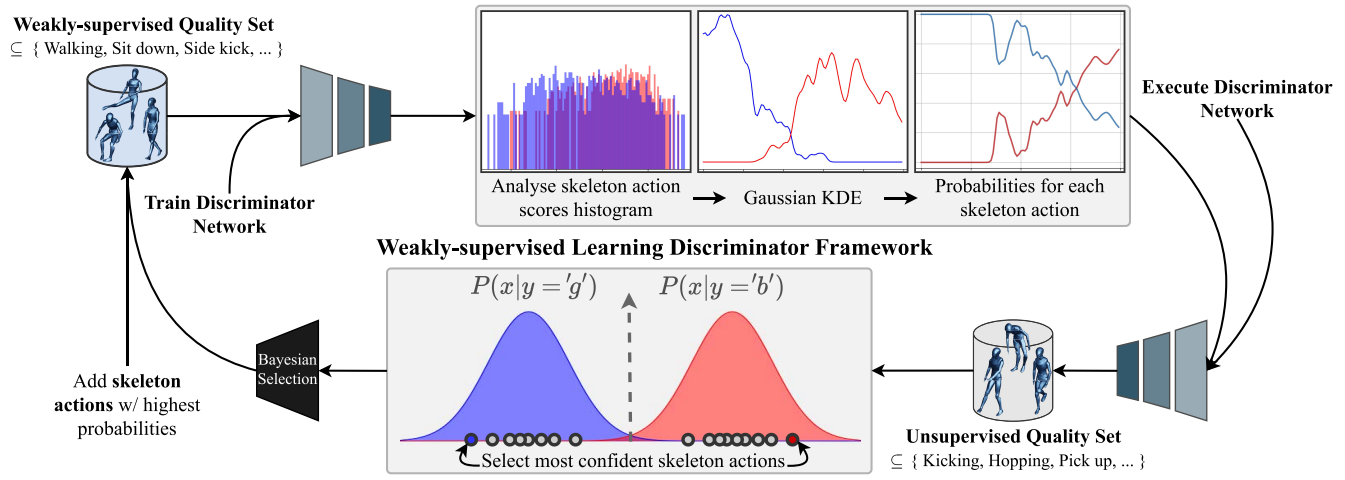


Fig. 2. **Illustrative Overview of the Weakly-supervised Learning Discriminator Strategy**. First, we generate a weakly-supervised set using existing generative methods like Kinetic-GAN [13], CSGN [55], or cGAN [36], and annotate a small set of generated samples as good or bad based on quality. Next, we train a Graph Convolutional Network to distinguish between well-executed and poorly executed skeletons and use its output scores to filter out unreliable instances from an unlabelled dataset. Then, we combine the discriminator with a Bayesian classifier to identify the most reliably classified samples, which are fed into an action recognition model. The selected samples are also used to train the discriminator in the next iteration, thereby improving the framework's overall performance.

of the corresponding samples, respectively. Additional metrics, such as the maximum mean discrepancy (MMD) [47], are also commonly employed in human action synthesis for comparing features of two sets of data. In addition to their dependency on real data for comparison, these metrics also have a limitation in their ability to provide detailed, sample-level quality estimation. As distribution-based metrics, they cannot capture detailed information at the individual sample level.

This paper introduces an alternative and complementary evaluation metric. It aims to not only analyze the individual quality of samples but also move towards assessing the contextual quality of generated samples and their intended purpose. An additional advantage of this metric lies in its independence from real data, saving both labeling time and enabling direct application to generated data for low-computational-cost data augmentation.

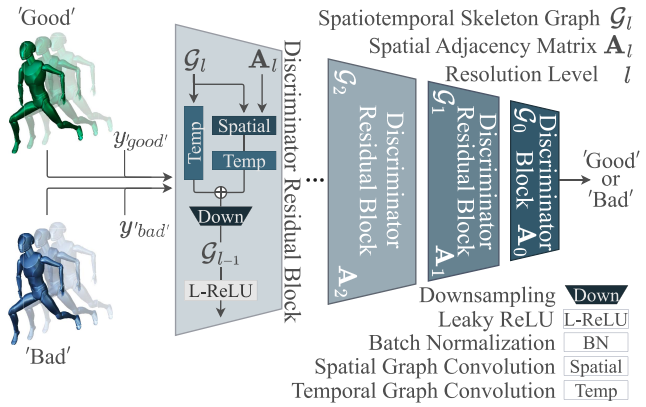


Fig. 3. **Discriminator**. The discriminator takes as input a skeleton graph sequence $\mathcal{G}_l$, along with its embedded class (achieved through channel-wise concatenation), and learns to discriminate by coarsening spatially and temporally from level l to 0. This downsampling process results in a more efficient and effective discrimination process, enabling our model to better filter out low-quality samples.

III. Iterative Weakly-Supervised Learning Generative Framework

A. Weakly-Supervised Discriminator Network

Graph Convolutional Networks (GCNs) allow the extraction of embedded patterns over spatial and temporal axes of a skeleton sequence. This is accomplished by generalizing convolution operations (mainly used for images) to graphs. Based on this approach, we use a Spatiotemporal GCN discriminator as our quality network to extract quality patterns of a skeleton action (Fig. 3).

Graph Convolutional Networks (GCNs) leverage a spatiotemporal graph $\mathcal{G}_l = (\mathcal{V}_l, \mathcal{E}_l)$ to represent skeleton data with N_l joints and T_l frames, where L is the number of levels of the skeleton graph resolution and $l = \{1, \ldots, L\}$. In this context, the feature map of the skeleton sequence can be represented by $\mathbf{X}_l \in \mathbb{R}^{N_l \times T_l \times C}$, where C is the number of channels that represent joint coordinates at resolution level l. A GCN comprises both spatial and temporal graph convolutions. In the spatial dimension, the intra-body joint connections are typically defined using an adjacency matrix $\mathbf{A}_l \in \{0,1\}^{N_l \times N_l}$ and the corresponding identity matrix $\mathbf{I}_l$. These matrices are used to regulate the receptive fields of the convolution. To enable the construction of convolution operations, a partitioning strategy is defined to represent the neighbor set of each joint due to the high-level formulation of the problem. In this context, $\mathbf{A}_l$ and $\mathbf{I}_l$ are dismantled into three partitions p, based on the spatial configuration proposed by [56]. Thus, we have $\mathbf{A}_l + \mathbf{I}_l = \sum_p \mathbf{A}_{l_p}$. Given that multiple graph convolutional layers are used, each layer may contain different levels of semantic information [44], [56]. The problem with this is that simply using $\mathbf{A}_l$ forces the same pre-defined spatial weight for all layers. Therefore, we introduce a learnable weight matrix $\mathbf{M}_l \in \mathbb{R}^{N_l \times N_l}$, initialized as an all-one matrix. For each layer of

the generator and discriminator. This enables us to adaptively optimize the spatial weight configuration of $\mathbf{A}_l$, and compute the graph convolution as follows:

$$\mathcal{S}(\mathbf{X}_l) = \sum_{i=1}^{p} \Lambda_{l_i}^{-\frac{1}{2}} \left(\mathbf{A}_{l_i} \odot \mathbf{M}_l\right) \Lambda_{l_i}^{-\frac{1}{2}} \mathbf{X}_l \mathbf{W}_{l_i}, \tag{1}$$

where the normalization of the adjacency matrix $\mathbf{A}_{l_p}$ is achieved through the degree matrix $\Lambda_{l_p}^{ii} = \sum_j (\mathbf{A}_{l_p}^{ij})$, which represents the summation of edges attached to each joint node. The weight vectors for each partition group p from resolution level l are stacked and represented by $\mathbf{W}_{l_p}$.

Over the temporal axis, we propose the use of consecutive frames as consecutive skeletons to leverage one-dimensional kernels in temporal graph convolution. This operation is applied after the spatial graph convolution. Finally, the spatiotemporal graph convolution is computed by convolving the positional features joint-wise, as $\mathcal{T}\big(\mathcal{S}(\mathbf{X}_l)\big) = \mathcal{S}(\mathbf{X}_l) * \mathbf{w}_l$, where $\mathbf{w}_l \in \mathbb{R}^{1\times t\times C}$ is the temporal kernel at resolution l with t as the number of frames to be convolved in the kernel. By computing the spatiotemporal graph convolution at every level l of the discriminator, we can inherently capture the temporal evolution of joint positions and connections over time, thus producing a sequential spatiotemporal representation of the action. Extracting spatiotemporal features from a complete action $\mathbb{X} = \langle x_1, x_2, \ldots, x_n \rangle$ enables capturing quality patterns without losing any spatial or temporal dependencies. These features allow a direct analysis of the spatiotemporal evolution of joint positions and connections over time and guide the learning process of the discriminator with a binary cross-entropy loss.

B. Bayesian Classifier

The proposed weakly-supervised discriminator network adopts a Multiple Instance Learning (MIL) paradigm, treating input skeleton sequences as bags labeled in a binary fashion. Here, good bags represent well-structured skeletons executing accurate actions, while bad bags denote poorly structured and executed ones. To delve into the theoretical underpinnings, our approach aligns with weakly supervised learning principles, where bag-level annotations guide the discriminator network. Furthermore, self-supervised learning elements are inherently present, as the iterative generation strategy refines the model through its own generated data, contributing to a more comprehensive understanding of well-performed skeleton actions. In essence, our method leverages both weakly supervised learning and self-supervised learning aspects for more effective action synthesis.

After the initial generation of learning using a small set of synthesized samples, Bayesian classifiers can be used to obtain a degree of belief (Fig. 4) for each classified instance in the unsupervised set. The key idea behind this approach is to select only instances with an extremely high belief for the next generation of the learning set. This is accomplished in a self-supervised manner for both the weakly-supervised discriminator network (WSDN) and the action recognition model.

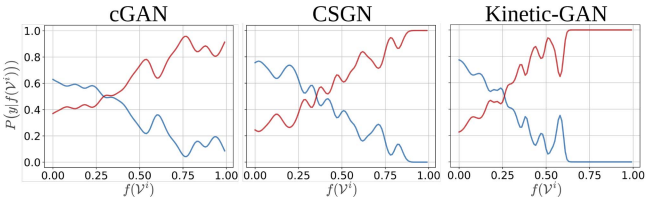


Fig. 4. **Comparison between the posteriors** $P\big(y|f(\mathcal{V}^i)\big)$ obtained by the Bayesian classifier across each generative baseline.

The Bayesian classifier is embedded in our method's framework to achieve self-supervision by filtering out the actions that should be used for training the action recognition model accordingly to the received scores $f(\mathcal{G})$ produced by the discriminator network:

$$P\big(y|f(\mathcal{G}^i)\big) = \frac{P\big(f(\mathcal{G}^i)|y\big)P(y)}{P\big(f(\mathcal{G}^i)\big)} \tag{2}$$

where $y \in \{'g', 'b'\}$ represent the *good/bad* quality classes. To approximate the conditional densities $P\big(f(\mathcal{G}^i)|y\big)$, a Gaussian kernel density estimator with Scott's rule [41] for bandwidth selection was used. During our experiments and annotations, we noticed a balance between good and bad instances with Kinetic-GAN [13] and CSGN [55]. However, to address the substantial imbalance between the number of bad instances and good ones over cGAN [36], the priors were empirically adjusted to $P('g') = P('b') = 0.5$.

Formally, the rule for selecting the i-th skeleton action for the next generation of the action recognition model learning data is:

$$\boldsymbol{\mathcal{G}}^{(t+1)} \stackrel{\text{def}}{=} \left\{\mathcal{G}^i \iff P\big(y|f(\mathcal{G}^i)\big) \geq \tau_1\right\}, y \in \left\{'g', 'b'\right\}, \tag{3}$$

i.e., an unsupervised skeleton action's prediction score is selected if its posterior probability for either the $'b'$ or $'g'$ class is above a certain threshold.

C. Conditional Generative Adaptation

Generating specific actions is crucial for human action synthesis. However, previous generative models for synthesizing skeleton actions were not designed to generate controllable motions. To address this limitation, we adapted state-of-the-art approaches, including CSGN and a skeleton-based GAN, to enable greater control over the action generation process. Specifically, in both generators, we incorporated the embedded class representation y into the input noise z, such that the generative process is aware of which action is being generated [13]. Meanwhile, the discriminators are fed with the channel-wise concatenation of the skeleton and the embedded class representation y. In our approach, we applied the WGAN-GP [19] objective formulation, which is conditioned as:

$$\min_G \max_D \overbrace{\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_r}\big[D(\boldsymbol{x}|\boldsymbol{y})\big] - \mathbb{E}_{\tilde{\boldsymbol{x}}\sim\mathbb{P}_g}\big[D(\tilde{\boldsymbol{x}}|\boldsymbol{y})\big]}^{\text{Discriminator loss}} + \lambda \overbrace{\mathbb{E}_{\hat{\boldsymbol{x}}\sim\mathbb{P}_{\hat{\boldsymbol{x}}}}\Big[\big(\big\|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}}|\boldsymbol{y})\big\|_2 - 1\big)^2\Big]}^{\text{Gradient penalty}}, \tag{4}$$

where $\mathbb{P}_r$ is the data distribution and $\mathbb{P}_g$ is the model distribution implicitly defined by $\tilde{\boldsymbol{x}} = G(\boldsymbol{z}, \boldsymbol{y}), \boldsymbol{z} \sim p(\boldsymbol{z})$ (the input $\boldsymbol{z}$ is sampled from a noise distribution p, which is then concatenated with the embedded action class representation $\boldsymbol{y}$). $\mathbb{P}_{\hat{\boldsymbol{x}}}$ is sampled uniformly along straight lines, using linear interpolation, between pairs of points sampled from the data distribution $\mathbb{P}_r$ and generator distribution $\mathbb{P}_g$. The loss weight λ for gradient penalty is set to 10 in all experiments.

IV. Assessing Human Body Quality

The main idea behind synthetic human actions relies on capturing the dynamics of the human body's skeleton. These are rooted in the understanding of human body physics. However, existing methods for generating human actions primarily evaluate their models using Inception Score (IS) and Frechet Inception Distance (FID). Unlike the conventional approaches of using Inception-v3 trained with a large dataset of millions of images (ImageNet), current studies employ an action classifier trained on a smaller set of several thousand human actions. Despite the good correlations achieved by such evaluation metrics, several works have reported multiple flaws [2], [3], [37], such as misleading results when applying the classifier on datasets other than ImageNet, also leading biased gradients when evaluating a reduced number of samples also leading to misleading results and even single quality estimation.

In this section, we derive a novel notion of assessing synthetic human actions based on their physical attributes, including JERK and JITTER. This assessment method is applicable to unlabeled data and can be conducted independently of a real distribution. In Section IV-B, we further conduct a thorough analysis to examine how our evaluation correlates with the quality of outputs generated by various state-of-the-art human action generative models compared to traditional evaluation metrics. Additionally, we validate our approach by comparing it with human-annotated synthetic samples.

A. Assessing Skeleton Generative Models via Kinematics

An action performed by a human skeleton involves the coordinated movement of multiple joints, forming a structured and dynamic representation in both the spatial and temporal dimensions. The inherent nature of skeleton-based data facilitates its analysis when compared to images and videos. This advantage allows us to leverage physics-based metrics specifically designed for this data type. Subsequently, we assess the attributes of synthetic samples by analyzing their velocity, acceleration, JERK, and JITTER. These metrics provide valuable insights into the dynamic aspects of the generated actions, allowing for a comprehensive assessment of their realism and quality. In the context of skeleton data, we measure how quickly a skeleton $\mathcal{G}$ moves over time (velocity) through $\frac{1}{N \cdot T} \sum_{n=1}^{N} \sum_{t=1}^{T} \| \frac{dp_n(t)}{dt} \|$ and how the velocity of a skeleton changes over time (acceleration) through $\frac{1}{N \cdot T} \sum_{n=1}^{N} \sum_{t=1}^{T} \| \frac{d^2 p_n(t)}{dt^2} \|$, where $p_n(t)$ represents the n-th joint position at time t.

As is often observed in image modelling [4], [17], the pursuit of higher quality in generated samples can often result in the appearance of undesirable artifacts. This phenomenon, which is well-established in the image-based field [24], [52], holds true for synthetic human actions as well. Artifacts frequently arise due to the use of normalization methods within the generator, which enhance training stability by mitigating covariate shifts. However, the normalization of feature maps attenuates the information regarding the magnitude of individual features. It is assumed that the generator amplifies these magnitudes, which goes unnoticed by the discriminator during the normalization process [13], [24], [52].

This leads to the generation of synthetic actions that exhibit unnatural human dynamics, including irregular fluctuations, unstable acceleration, oscillations, and trembling. These anomalies can be discerned by analyzing the fundamental principles of human physics. Therefore, we further investigate the JERK of a skeleton action, which measures the rate of acceleration changes over time. By quantifying the smoothness or abruptness of these changes within the skeleton, JERK provides insights into the overall dynamics of the action, which is defined as:

$$jerk_{\text{overall}}(\mathcal{G}) = \frac{1}{T \cdot N} \sum_{t=1}^{T} \sum_{n=1}^{N} \left\| \frac{d^3 p_n(t)}{dt^3} \right\| \tag{5}$$

Furthermore, we also analyse the occurrence rate of position or orientation fluctuations in the skeleton's joints $\mathcal{G}$ over time. This phenomenon, known as jitter, quantifies the level of instability or variability exhibited by the joints during the action, which is defined as:

$$jitter_{\text{overall}}(\mathcal{G}) = \frac{1}{T \cdot N} \sum_{t=1}^{T} \sum_{n=1}^{N} \| p_n(t) - p_n(t-1) \| \tag{6}$$

By applying these metrics, we can assess the smoothness of transitions between various poses or movements within the skeleton. This is achieved by analysing the continuity and regularity of joint trajectories and motion sequences, allowing us to gauge the natural and seamless flow of the skeleton's movements. Fig. 5 depicts two generated actions of a person jumping, where the lower sample (blue, generated by Kinetic-GAN [13]) exhibits significantly smoother motion compared to the upper sample (red, generated by cGAN [36]). While both generative methods effectively model the human structure, there is a notable disparity in the quality of human dynamics they produce.

B. Sample Quality Assessment and Analysis

In recent skeleton generative models, generating high-quality examples among a set of outputs still remains challenging. Existing state-of-the-art evaluation metrics primarily compare samples from the real and generated distributions, resulting in a limited emphasis on assessing the individual quality of samples [1]. Currently, the selection of the best-generated samples still relies on human confirmation, indicating a continued dependence on human judgment in the process.

Hence, we adopt an alternative approach to evaluate generative models. Rather than assessing the generative distribution

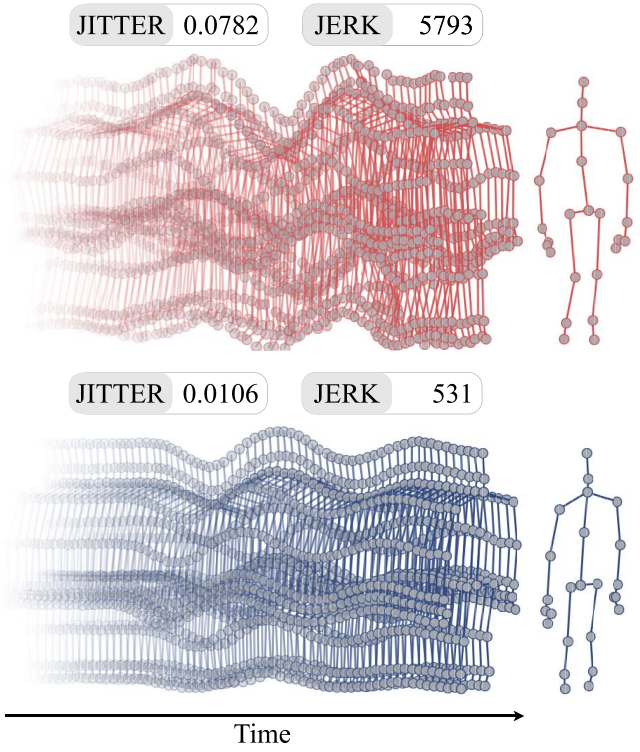


Fig. 5. **Comparing skeleton dynamics smoothness**. The difference between a "jump up" action poorly executed by cGAN [36] (red skeleton action), and a well executed one from Kientic-GAN [13] (blue skeleton action).

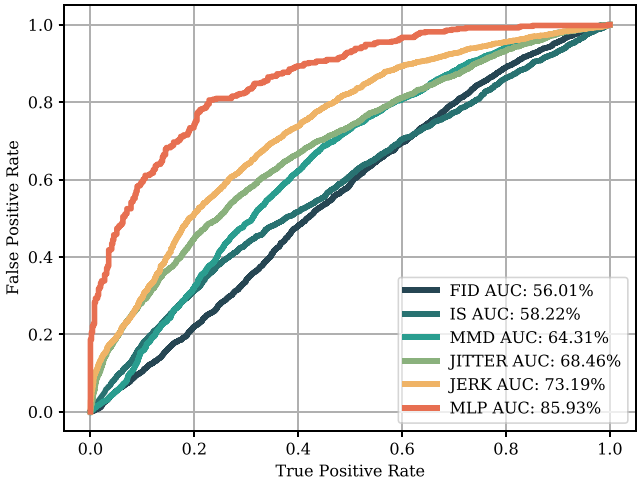


Fig. 6. **Individual Quality Metrics Performance**. ROC curves comparison obtained from 10,000 annotated samples with 5,000 high-quality actions and 5,000 low-quality actions.

based on collective measures such as the likelihood or statistical divergence, we individually classify each sample as high or low quality. In order to validate the effectiveness of our physics-based metrics in contrast to previous state-of-the-art evaluation metrics, we conducted an experiment utilising a set of 10,000 annotated samples from Kinetic-GAN's synthetic actions [13]. From this set, we obtained 5,000 well-performed skeleton actions (annotated as high quality) and an additional 5,000 samples annotated as low quality. Fig. 6 illustrates a comparison of the ROC curves for current state-of-the-art evaluation metrics commonly employed in skeleton-based

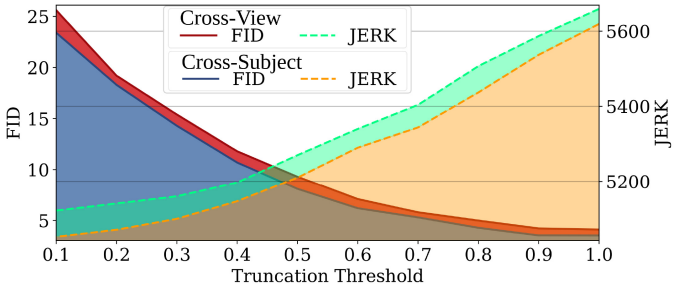


Fig. 7. **"Truncation trick" effect analysis**. The FID and JERK (both lower the better) evolution w.r.t. truncated sampling distributions of the latent space with both benchmarks of NTU RGB+D [42].

action synthesis, such as FID, IS, and the Maximum Mean Discrepancy (MMD), along with the physics-based metrics, JERK and JITTER. Additionally, we trained a multilayer perceptron using an additional set of 1,000 samples comprising both low and high-quality actions. The features used for training were extracted from a conventional spatiotemporal graph convolution [56]. Given that FID, IS, and MMD are distance metrics comparing two distributions, we applied a bootstrapping-like strategy by executing each metric 1,000 times for each sample. We then calculated the average metric value to assess the quality of each sample.

Furthermore, an analysis of the distribution of training data reveals that regions with low density are noticeably underrepresented, making it challenging for the generator to learn how to accurately model such areas. As established in prior studies [4], [23], [27], [35], the quality of generated samples can be enhanced by utilising truncated or shrunken sampling distributions. Despite potential losses in variation, we adopt a similar approach to analyse the behaviour of physics-based metrics using Kinetic-GAN mapping network [13]. During the inference phase, we calculate the scaling factor for the deviation of a given intermediate latent point $\boldsymbol{w}$ from the center of mass of $\boldsymbol{\mathcal{W}}$ using the following expression:

$$\boldsymbol{w}' = \mathbb{E}_{z\sim\mathbb{P}_z}\big[f(z)\big] + \psi\big(\boldsymbol{w} - \mathbb{E}_{z\sim\mathbb{P}_z}\big[f(z)\big]\big), \tag{7}$$

where $\psi \leq 1$, $f(\cdot)$ denotes the projected latent space, and $\mathbb{P}_z$ is the latent space distribution from 1,000 points. As shown in Fig. 7, it is evident that the quality of generation can be enhanced by observing the JERK of truncated synthetic samples as the shrinking factor ψ increases, which is consistent with the findings of prior studies [4], [23], [24]. In contrast to the FID, this observation further validates our findings regarding the individual quality of samples. Specifically, we observe that as the threshold ψ decreases to ≤ 0.9, the variation begins to diminish, consequently leading to a significant increase in FID scores. It is worth noting that conducting an experiment that combines both metrics can serve as a valuable starting point for selecting the best-quality samples and parameters.

V. Experiments and Discussion

A. Datasets and Experimental Settings

NTU RGB+D [42]. The dataset comprises 56,880 videos that belong to 60 action classes. For each sample, 3D skeleton data captured from 25 joints of 40 volunteers is provided.

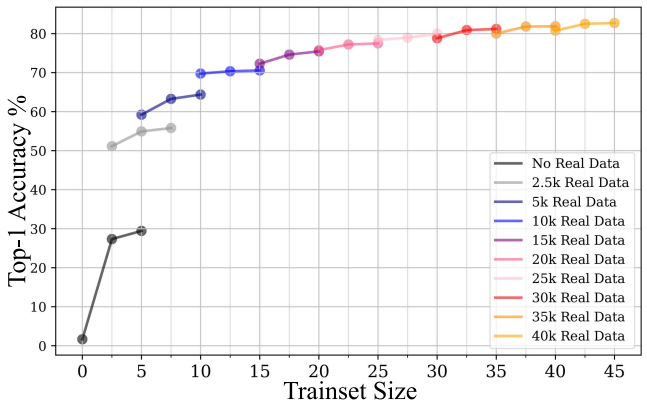


Fig. 8. **Accuracy evolution** on NTU RGB+D [42]. The first result of each line corresponds to the accuracy obtained using only real data, followed by the accuracy obtained after adding 2,500 and 5,000 synthetic samples in each experiment. The mean accuracy over five experimental runs (including real data) is reported.

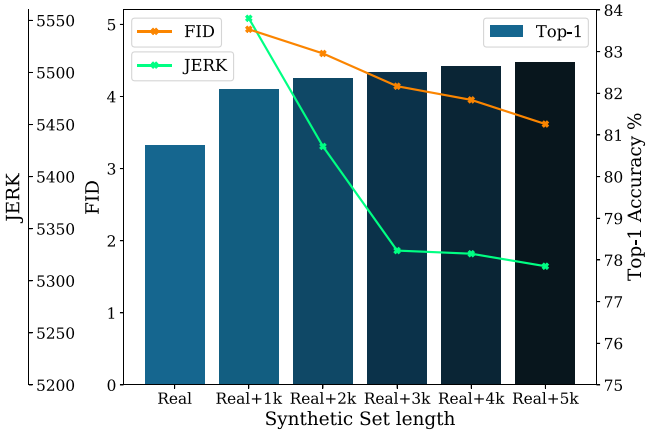


Fig. 9. **JERK vs. FID vs. Accuracy** on NTU RGB+D [42]. The graph shows the relationship between accuracy, FID and JERK metrics (both lower the better) for each increasing set size, except for the case where only real data is used.

The authors suggest two benchmarks for evaluating models: 1) cross-subject, where models are tested on other subjects than the ones they were trained on, and 2) cross-view, where models are tested on different camera views than the ones they were trained on..

NTU-120 RGB+D [30]. The dataset is an expanded iteration of its previous version, encompassing 114,480 video samples that are categorized into 120 action classes. The videos were captured using three camera views in 32 distinct setups, with data collected from 106 volunteers and 25 body joints per skeleton. The authors suggested two benchmarks for evaluation: 1) cross-subject, where models are tested on other subjects than the ones they were trained on, and 2) cross-setup, where models are tested on different setups than the ones they were trained on.

B. Performance Evolution

To validate the effectiveness of combining synthetic skeleton actions with real training data, we conducted an experiment using 10,000 samples from Kinetic-GAN's synthetic data [13]. From these, we obtained 5 thousand well-performed skeleton actions and added them to multiple training sets of varying sizes to observe their impact on accuracy. The accuracy evolution was plotted, starting from using only synthetic data with no real samples included, up to 40,000 real skeleton actions combined with 5,000 synthetic samples. Fig. 8 shows the results of this experiment using spatiotemporal graph convolutional network (ST-GCN) as the recognition model, where the same 5,000 synthetic samples are used across all training sets. The same real samples are also included in each subsequent training set.

Based on our experiment using Kinetic-GAN's synthetic data [13], we observed that the improvement in action recognition performance is greater when the real training set size is smaller. However, even with a larger training set of 40 thousand real skeleton samples, we still achieved a significant performance gain of 2.8% (3.4% p.p.). This suggests that synthetic data can be a valuable addition to real training data, even when the real data set is relatively large.

TABLE I
EVALUATING DIFFERENT GENERATOR DESIGNS. THE FID, MMD AND JERK SCORES (LOWER IS BETTER) BETWEEN REAL AND SYNTHETIC SAMPLES GENERATED

	Method	FID	MMD	JERK	FID	MMD	JERK
		Cross-Subject			*Cross-View*		
NTU RGB+D	$\not c$ GAN [37]	42.577	1.543	18493	47.936	1.789	22018
	Adapted cGAN	27.480	0.919	12732	31.875	0.993	17820
	WSDN-cGAN	**24.356**	**0.885**	**10821**	**28.454**	**0.934**	**15733**
	$\not c$ CSGN [56]	6.030	0.873	7543	7.114	0.910	8909
	Adapted CSGN	5.866	0.834	7004	6.790	0.873	8192
	WSDN-CSGN	**5.351**	**0.841**	**6359**	**6.133**	**0.844**	**7073**
	$\not c$ Kinetic-GAN	3.813	0.828	6130	4.593	0.849	6465
	Kinetic-GAN	3.618	0.772	5821	4.235	0.824	6009
	WSDN-Kinetic-GAN	**3.450**	**0.743**	**5103**	**3.865**	**0.799**	**5350**

$\not c$ No action conditioning.

Furthermore, we analyse the recognition accuracy in relation to the FID and JERK metrics. Fig. 9 depicts that the closer the convergence is (lower the FID and JERK values), the better the accuracy as more diverse characteristics of the skeletons (FID) are generated in the synthetic data, while containing more well-performed (JERK) action samples.

C. Ablation Study and Weakly-Supervised Experiments

In this section, we demonstrate the experimental improvements achieved by our conditional adaptation of state-of-the-art generative methods and the effectiveness of our weakly-supervised discriminator network (WSDN) in enhancing the quality of their generated data.

Table I depicts a comparison of FID, MMD and JERK for various generator architectures on the NTU RGB+D benchmark dataset. We evaluate each distribution under the respective settings of each method. Our first observation is the significant quality improvement brought about by the architecture of previous generative models, specifically the adapted CGAN and adapted CSGN, with conditional training that incorporates class-embedded information into their GAN architecture.

As our work is based on the premise that a set of unsupervised data is available, we limited the number of annotated

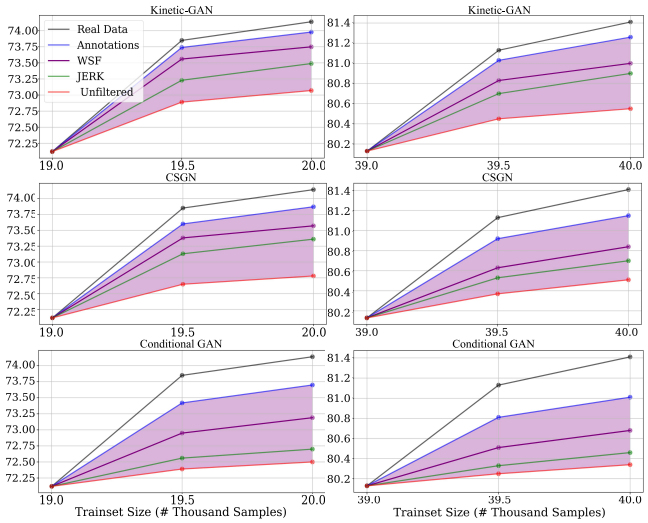


Fig. 10. **Accuracy Evolution**. The graph displays the accuracy evolution of synthetic data for each baseline compared to using real data, starting from 19,000 and 39,000 real samples. WSF stands for our weakly-supervised framework.

samples to 1,000 from each generative model – Kinetic-GAN, Adapted CSGN, and Adapted cGAN - and included 100,000 synthetic unsupervised samples from the respective baseline. In practice, each experiment starts rigorously with the same amount of annotated labels. To further enhance the quality of the generated samples from all three baselines, we employ an iterative selection process through our WSDN framework, which further exhibits quality improvement compared to the original samples from the adapted architectures. This demonstrates the effectiveness of our proposed WSDN framework in enhancing the quality of the generated samples.

To evaluate the performance evolution, we conducted a five-fold experiment, starting with 19,000 and 39,000 real samples. We compared the evolution of different sample sets, including 1,000 real samples, 1,000 annotated synthetic samples, 1,000 samples generated by our WSDN framework, the top 1,000 samples filtered by JERK, and 1,000 randomly-samples synthetic samples. These sets were individually compared to their respective baselines. While using generated data did not surpass the performance of real data, our results demonstrate a significant improvement without the need for fine-grained annotation (Fig. 10). Specifically, we can improve recognition accuracy up to 1.23% using only 1,000 annotated synthetic samples, 0.97% with only 1,000 WSDN-generated samples, and 0.86% using just the top-1,000 samples from JERK. The pink region indicates the potential for improvement that our WSDN approach can achieve, which is encouraging as it is closer to the performance of the synthetic annotated data compared to unfiltered/random samples. Even with randomly generated samples from the conditional training adaptation, our method achieved a slight improvement in the results, which can also be positively regarded. This demonstrates the effectiveness of our approach and highlights the potential for further improvement with more refined generative models and iterative training. Table II and III showcases the performance of our proposed framework on two benchmark datasets, namely NTU RGB+D [42] and NTU-120 RGB+D [30]. For each synthetic baseline, we rigorously used only 5,000 samples generated by our method, in addition to the full set of real (supervised) data.

TABLE II
SEMI-SUPERVISED LEARNING EVALUATION ON NTU RGB+D [42]. TOP-1 ACCURACY REPORTED FROM SMALL PERCENTAGES OF TRAINING DATA. THE FIRST ROW OF EACH BACKBONE NETWORK REPRESENTS THE RESULTS OF ONLY USING THE SUPERVISED DATA

			Real Data			
	Arch	**Method**	**50%**	**100%**	**50%**	**100%**
NTU RGB-D			*Cross-Subject*		*Cross-View*	
	ST-GCN	Supervised [57]	74.14	81.51	78.35	88.34
		SKL-VAE	74.19	81.52	78.41	88.52
		cGAN	74.25	81.67	78.64	88.98
		SA-GCN	74.90	81.83	79.01	89.12
		CSGN	75.42	82.24	79.75	89.95
		Kinetic-GAN	**76.45**	**83.21**	**80.59**	**90.71**
	As-GCN	Supervised [30]	78.55	86.96	79.14	94.15
		SKL-VAE	78.60	86.04	79.22	94.17
		cGAN	78.67	87.34	79.44	94.29
		SA-GCN	78.93	87.55	80.13	94.47
		CSGN	80.32	88.53	81.26	95.33
		Kinetic-GAN	**80.99**	**89.42**	**81.82**	**95.89**
	2s-AGCN	Supervised [45]	79.44	88.51	80.82	95.12
		SKL-VAE	79.52	88.63	80.89	95.20
		cGAN	79.56	88.77	80.95	95.27
		SA-GCN	80.00	89.04	81.33	95.75
		CSGN	80.74	89.65	81.83	96.15
		Kinetic-GAN	**81.32**	**90.15**	**82.54**	**96.72**
	CTR-GCN	Supervised [7]	83.37	92.42	84.65	96.79
		SKL-VAE	83.38	92.77	84.70	96.83
		cGAN	83.41	93.01	84.79	96.90
		SA-GCN	83.83	93.21	85.06	97.20
		CSGN	84.29	93.40	85.54	97.55
		Kinetic-GAN	**84.80**	**93.98**	**86.77**	**98.01**
	Info-GCN	Supervised [9]	83.97	93.03	85.32	97.11
		SKL-VAE	83.99	93.11	85.39	97.14
		cGAN	84.03	93.53	85.44	97.20
		SA-GCN	84.41	93.87	85.92	97.44
		CSGN	84.95	94.44	86.29	97.89
		Kinetic-GAN	**85.36**	**94.98**	**87.23**	**98.39**

D. Distribution Complementarity

In addition to evaluating the overall quality of synthesized actions, we also explore how diversity across kinematic features can be achieved by combining real and generated data. The performance of a single action can vary from person to person, scenario to scenario, and even when performed by the same person twice. To account for such variability, we have identified several kinematic features that are relevant to the motion and movement of each skeleton. These features also include motion energy, which characterizes the temporal dynamics by capturing the amount of movement and how it changes over time within an action.

While generative models have the ability to create synthesized actions within the range of real data, the quality of the generated data can vary significantly depending on the architecture of the generator. To investigate this hypothesis, we utilize shape metric distributions, such as skewness and kurtosis, to directly compare the characteristics of real data with synthetic data. By measuring the asymmetry of the

TABLE III
SEMI-SUPERVISED LEARNING EVALUATION ON NTU-120 RGB+D [30]. TOP-1 ACCURACY REPORTED FROM SMALL PERCENTAGES OF TRAINING DATA. THE FIRST ROW OF EACH BACKBONE NETWORK REPRESENTS THE RESULTS OF ONLY USING THE SUPERVISED DATA

			Real Data			
	Arch	**Method**	**50%**	**100%**	**50%**	**100%**
			Cross-Subject		*Cross-Setup*	
NTU-120 RGB-D	**ST-GCN**	Supervised [57]	69.43	75.08	70.01	76.12
		SKL-VAE	69.49	75.23	70.12	76.28
		cGAN	69.51	75.66	70.30	76.65
		SA-GCN	69.97	75.71	70.77	76.99
		CSGN	71.20	76.78	71.42	77.54
		Kinetic-GAN	**72.32**	**77.65**	**72.51**	**78.11**
	As-GCN	Supervised [30]	73.54	78.37	73.78	78.92
		SKL-VAE	73.56	78.71	73.84	79.01
		cGAN	73.59	78.85	73.99	79.34
		SA-GCN	74.03	79.11	74.37	79.57
		CSGN	74.80	79.45	74.89	80.02
		Kinetic-GAN	**75.31**	**80.11**	**75.43**	**80.45**
	2s-AGCN	Supervised [45]	74.44	79.55	75.32	81.73
		SKL-VAE	74.50	79.59	75.38	81.77
		cGAN	74.58	79.67	75.45	81.83
		SA-GCN	75.01	80.20	75.74	82.28
		CSGN	75.55	80.54	76.23	82.69
		Kinetic-GAN	**76.13**	**80.93**	**76.96**	**83.22**
	CTR-GCN	Supervised [7]	82.31	88.85	83.22	90.63
		SKL-VAE	82.37	88.90	83.26	90.65
		cGAN	82.42	88.96	83.33	90.70
		SA-GCN	82.74	89.31	83.57	99.01
		CSGN	83.28	89.81	84.19	91.55
		Kinetic-GAN	**83.87**	**90.23**	**84.80**	**91.99**
	Info-GCN	Supervised [9]	82.41	89.82	83.56	91.16
		SKL-VAE	82.44	89.83	83.59	91.17
		cGAN	82.50	89.91	83.64	91.20
		SA-GCN	82.78	90.22	83.89	91.54
		CSGN	83.36	90.82	84.41	92.01
		Kinetic-GAN	**84.91**	**90.90**	**84.88**	**92.34**

distribution through skewness, the peakedness and flatness of a distribution through kurtosis, we can analyze how kinematic features vary between real and synthetic samples.

Given that the model's ability to generalize to unseen data is essential, we aimed to generate a diverse set of synthetic data to augment the real training set. By increasing its diversity, we hoped to create a more robust and generalizable model. In essence, our primary objective was to improve the model's ability to recognize actions that it had not encountered during training.

Fig. 11 demonstrates the degree of diversity achieved in the distribution of all four kinematic features while increasing both real and synthetic data. One key observation is that there are noticeable differences across all baselines, which is expected due to the inherent variation factors from the latent space [6], [14]. The second observation is the difference in diversity between Kinetic-GAN and CSGN generated skeletons when compared to real data. CSGN uses Gaussian processes to model the temporal relationships in a latent sequence, and as a result, from a generative perspective, there are more entangled latent factors with multiplicative

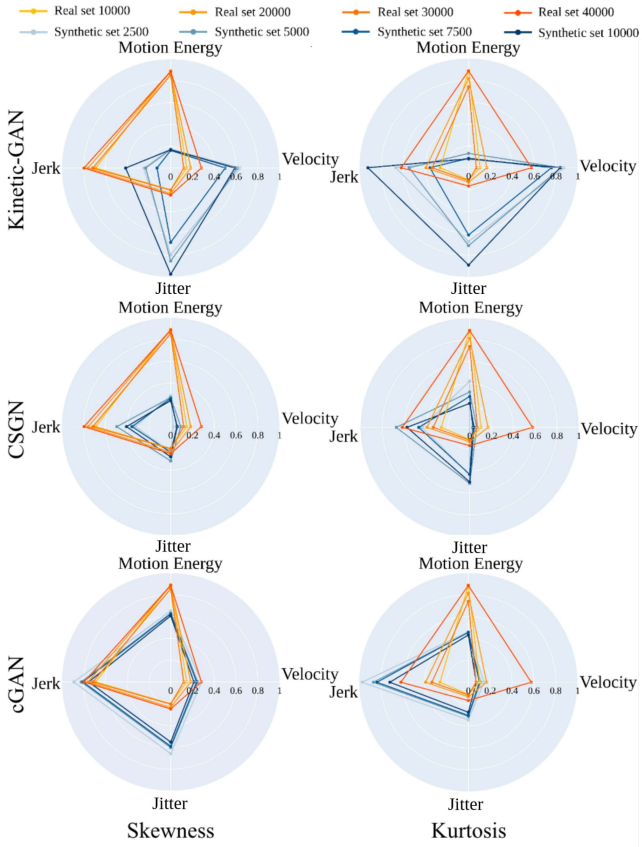


Fig. 11. **Skewness vs Kurtosis**. Comparison between the skewness and kurtosis of real and synthetic data used to train action recognition models.

interactions across the latent sequence. This has been previously demonstrated in image modeling [23], [24], [43]. On the other hand, Kinetic-GAN excels in terms of diversity by employing a mapping network that generates an intermediate latent space. This approach facilitates improved separation of various factors of variation within the data. Additionally, Kinetic-GAN includes a noise injection module which introduces stochastic variation into the generation process. These two factors contribute to the ability of Kinetic-GAN to produce more diverse samples when compared to CSGN and cGAN.

VI. CONCLUSION, LIMITATIONS AND FURTHER WORK

In this paper, we propose employing generative adversarial and graph convolutional networks for augmenting skeleton-based human action data. By generating synthetic data that captures both the structural and temporal aspects of the data, the proposed method outperformssupervised state-of-the-art action recognition and synthesis methods in terms of accuracy and diversity. Furthermore, this work introduces an alternative approach to quantitatively evaluate generative models by considering the context of skeleton data. While physics-based metrics are manually crafted and may introduce biases in physics characteristics due to the selected samples, they introduce a novel perspective that, when combined with current state-of-the-art metrics, enhances the comparison of models. We investigate the properties of both proposed approaches

on four real-world benchmarks (NTU RGB+D and NTU-120 RGB+D), advancing the state-of-the-art performance metrics by a significant margin. Furthermore, existing skeleton-based generative models remain confined to the data on which they are trained, posing challenges in generating novel motions and actions—also acknowledged as a limitation in this work. Nevertheless, the integration of physics-guided metrics and knowledge represents an active area of research aimed at addressing and overcoming such limitations.

REFERENCES

[1] A. M. Alaa, B. Van Breugel, E. Saveliev, and M. van der Schaar, "How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models," in *Proc. ICML*, 2022, pp. 1–17.

[2] S. Barratt and R. Sharma, "A note on the inception score," 2018, *arXiv:1801.01973*.

[3] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *Proc. ICLR*, 2018, pp. 1–36.

[4] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. ICLR*, 2019, pp. 1–35.

[5] H. Cai, C. Bai, Y.-W. Tai, and C.-K. Tang, "Deep video generation, prediction and completion of human action sequences," in *Proc. ECCV*, 2018, pp. 1–17.

[6] T. Qi Chen, X. Li, R. B. Grosse, and D. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," in *Proc. NeurIPS*, 2018, p. 31.

[7] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proc. ICCV*, 2021, pp. 1–10.

[8] Z. Chen, S. Li, B. Yang, Q. Li, and H. Liu, "Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition," in *Proc. AAAI*, 2021, pp. 1–10.

[9] H.-G. Chi, M. H. Ha, S. Chi, S. Wan Lee, Q. Huang, and K. Ramani, "InfoGCN: Representation learning for human skeleton-based action recognition," in *Proc. CVPR*, 2022, pp. 1–11.

[10] B. Chopin, N. Otberdout, M. Daoudi, and A. Bartolo, "3-D skeleton-based human motion prediction with manifold-aware GAN," *IEEE Trans. Biometr., Behavior, Identity Sci.*, vol. 5, no. 3, pp. 321–333, Jul. 2023.

[11] B. Degardin, V. Lopes, and H. Proença, "REGINA—Reasoning graph convolutional networks in human action recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 5442–5451, 2021.

[12] B. Degardin, V. Lopes, and H. Proença, "ATOM: Self-supervised human action recognition using atomic motion representation learning," *Image Vis. Comput.*, vol. 137, Sep. 2023, Art. no. 104750.

[13] B. Degardin, J. Neves, V. Lopes, J. Brito, E. Yaghoubi, and H. Proença, "Generative adversarial graph convolutional networks for human action synthesis," in *Proc. WACV*, 2022, pp. 1150–1159.

[14] G. Desjardins, A. Courville, and Y. Bengio, "Disentangling factors of variation via generative entangling," 2012, *arXiv:1210.5474*.

[15] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. CVPR*, 2015, pp. 1110–1118.

[16] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *Proc. CVPR*, 2015, pp. 1–9.

[17] L. Galteri, L. Seidenari, M. Bertini, and A. Del Bimbo, "Deep generative adversarial compression artifact removal," in *Proc. ICCV*, 2017, pp. 4826–4835.

[18] I. J. Goodfellow et al., "Generative adversarial networks," 2014, *arXiv:1406.2661*.

[19] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. NeurIPS*, 2017, pp. 1–11.

[20] D. Gupta, S. Maheshwari, S. S. Kalakonda, M. Vaidyula, and R. K. Sarvadevabhatla, "DSAG: A scalable deep framework for action-conditioned multi-actor full body motion synthesis," in *Proc. WACV*, 2023, pp. 1–9.

[21] I. Habibie, D. Holden, J. Schwarz, J. Yearsley, and T. Komura, "A recurrent variational autoencoder for human motion synthesis," in *Proc. BMVC*, 2017, pp. 1–12.

[22] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. NeurIPS*, 2017, pp. 1–12.

[23] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. CVPR*, 2019, pp. 1–10.

[24] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. CVPR*, 2020, pp. 1–10.

[25] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for Senjian 3-D action recognition," in *Proc. CVPR*, 2017, pp. 1–10.

[26] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Learning clip representations for skeleton-based 3-D action recognition," *IEEE Trans. Image Process.*, vol. 27, pp. 2842–2855, 2018.

[27] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Proc. NeurIPS*, 2018, pp. 1–10.

[28] J. N. Kundu, M. Gor, P. K. Uppala, and V. B. Radhakrishnan, "Unsupervised feature learning of human actions as trajectories in pose embedding manifold," in *Proc. WACV*, 2019, pp. 1–9.

[29] B. Li, Y. Dai, X. Cheng, H. Chen, Y. Lin, and M. He, "Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN," in *Proc. ICMEW*, 2017, pp. 601–604.

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

[31] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3-D human action recognition," in *Proc. ECCV*, 2016, pp. 816–833.

[32] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention LSTM networks for 3D action recognition," in *Proc. CVPR*, 2017, pp. 3671–3680.

[33] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, Aug. 2017.

[34] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, "Are GANs created equal? A large-scale study," in *Proc. NeurIPS*, 2018, pp. 1–10.

[35] M. Marchesi, "Megapixel size image creation using generative adversarial networks," 2017, *arXiv:1706.00082*.

[36] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[37] M. Rosca, B. Lakshminarayanan, D. Warde-Farley, and S. Mohamed, "Variational approaches for auto-encoding generative adversarial networks," 2017, *arXiv:1706.04987*.

[38] M. S. M. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly, "Assessing generative models via precision and recall," in *Proc. NeurIPS*, 2018, pp. 1–10.

[39] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. NeurIPS*, 2016, pp. 1–9.

[40] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe, "MeTRAbs: Metric-scale truncation-robust heatmaps for absolute 3-D human pose estimation," *IEEE Trans. Biometr., Behavior, Identity Sci.*, vol. 3, no. 1, pp. 16–30, Jan. 2021.

[41] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. Hoboken, NJ, USA: Wiley, 2015.

[42] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+ D: A large scale dataset for 3D human activity analysis," in *Proc. CVPR*, 2016, pp. 1–10.

[43] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in *Proc. CVPR*, 2020, pp. 1–10.

[44] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. CVPR*, 2019, pp. 1–10.

[45] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. AAAI*, 2017, pp. 1–8.

[46] T. S. Kim and A. Reiter, "Interpretable 3-D human action analysis with temporal convolutional networks," in *Proc. CVPR Workshops*, 2017, pp. 1–8.

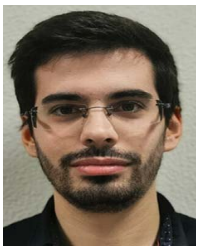
[47] D. J. Sutherland et al., "Generative models and model criticism via optimized maximum mean discrepancy," in *Proc. ICLR*, 2021, pp. 1–11.

[48] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. CVPR*, 2016, pp. 1–10.

[49] J. Tu, H. Liu, F. Meng, M. Liu, and R. Ding, "Spatial-temporal data augmentation based on LSTM autoencoder network for skeleton-based human action recognition," in *Proc. ICIP*, 2018, pp. 3478–3482.

[50] N. Wichers, R. Villegas, D. Erhan, and H. Lee, "Hierarchical long-term video prediction without supervision," in *Proc. ICML*, 2018, pp. 1–9.
[51] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *Proc. CVPR*, 2017, pp. 1–10.
[52] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. ECCV Workshops*, 2018, pp. 63–79.
[53] Z. Wang et al., "Learning diverse stochastic human-action generators by learning smooth latent transitions," in *Proc. AAAI*, 2020, pp. 1–8.
[54] B. Xu, X. Shu, and Y. Song, "X-invariant contrastive augmentation and representation learning for semi-supervised skeleton-based action recognition," *IEEE Trans. Image Process.*, vol. 31, pp. 3852–3867, 2022.
[55] S. Yan, Z. Li, Y. Xiong, H. Yan, and D. Lin, "Convolutional sequence generation for skeleton-based action synthesis," in *Proc. CVPR*, 2019, pp. 1–9.
[56] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI*, 2018, pp. 1–10.
[57] W. Yang, T. Lyons, H. Ni, C. Schmid, and L. Jin, "Developing the path signature methodology and its application to landmark-based human action recognition," in *Stochastic Analysis, Filtering, and Stochastic Optimization: A Commemorative Volume to Honor Mark HA Davis's Contributions*. Cham, Switzerland: Springer, 2022.
[58] P. Yu, Y. Zhao, C. Li, J. Yuan, and C. Chen, "Structure-aware human-action generation," in *Proc. ECCV*, 2020, pp. 1–17.
[59] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *Proc. ICCV*, 2017, pp. 1–10.
[60] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *Proc. CVPR*, 2020, pp. 1–10.
[61] Y. Zhou, Z. Li, S. Xiao, C. He, Z. Huang, and H. Li, "Auto-conditioned recurrent networks for extended complex human motion synthesis," in *Proc. ICLR*, 2018, pp. 1–13.

Bruno Degardin received the B.Sc. and M.Sc. degrees from the Universidade da Beira Interior (UBI) in 2018 and 2020, respectively, where he is currently pursuing the Ph.D. degree with his research interest focused on computer vision, emphasising biometrics, human behaviour analysis, and video understanding. In the past, he has worked on research projects, such as BIODI, CovidSight, and PAPSE-UBI. Since 2020, he has been a Teaching Assistant with UBI and a Co-Founder & the CTO with DeepNeuronic (Computer Vision solutions Startup). He received several awards, such as Best M.Sc. Thesis Fraunhofer Portugal Challenge (2020, 3rd place) and research of practical utility (1st place, Biomedical World Hackathon 2020, 1st place, UBIMedical Health Cup 2020, honourable mentions from APDC, and winner of WSA Portugal Awards).



Vasco Lopes received the B.Sc., M.Sc., and Ph.D. degrees in computer science and engineering in 2017, 2019, and 2023, respectively. He is currently an Invited Professor with the University of Beira Interior and since 2020, he has been the CEO and a Co-Founder of DeepNeuronic, a company that develops computer vision solutions to automate daily processes. In 2021, he collaborated with Huawei R&D Center, Paris, as a Research Assistant, where he developed Neural Architecture Search methods for network reliability, and as a Researcher with Google Research, Zurich, in 2022. He has worked on research projects, such as uPATO, RobotChain, INDTech 4.0, and CovidSight. He received several awards, such as the APRP Best M.Sc. Dissertation in Pattern Recognition Award in 2019 and the ICANN21 "1st Springer & ENNS Best Paper Award."

Hugo Proença (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in 2001, 2004, and 2007, respectively. He is currently a Full Professor with the Department of Computer Science, University of Beira Interior, Portugal, that has been researching mainly about biometrics and visualsurveillance. He was the Coordinating Editor of the IEEE BIOMETRICS COUNCIL NEWSLETTER and the Area Editor (ocular biometrics) of the IEEE BIOMETRICS COMPENDIUM journal. He is a member of the Editorial Board of the *Image and Vision Computing*, IEEE ACCESS, and *International Journal of Biometrics*. Also, he served as a Guest Editor of Special Issue of the *Pattern Recognition Letters*, *Image and Vision Computing*, and *Signal, and Image and Video Processing journals*.