Signal Processing: Image Communication 142 (2026) 117469

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

NFlowAD: A normalizing flow model for anomaly detection in human motion animations☆

Mahamat Issa Choueb [a], Praveen Kumar Sekharamantry [a,b,*], Giulia Martinelli [a], Francesco De Natale [a], Nicola Conci [a]

[a] DISI, University of Trento, 38123, Trento, Italy
[b] GSCSE, GITAM University, Visakhapatnam, 530045, Andhra Pradesh, India

ARTICLE INFO

Keywords:
Anomaly detection
Human motion
Self-supervised learning
Normalizing flows

ABSTRACT

Anomaly detection has been extensively investigated in numerous application areas. Hand-crafted rules have gradually given way to supervised classification techniques, which frequently rely on a small number of anomaly labels and related architectures. When it comes to human motion, abnormalities emerge at a fine-grained temporal or joint level rather than over a whole video sequence.

This study introduces NFlowAD, a self-supervised system that analyzes body joints to detect irregularities in human motion. It blends normalizing flows with masked motion modeling to describe normal motion data without the need for anomaly labels. Inference uses both reconstruction mistakes and flow-based likelihoods to detect anomalies. The validation pipeline on various state-of-the-art datasets demonstrates NFlowAD's efficiency in recognizing, locating, and analyzing anomalous motion sequences, while maintaining robust detection and interpretability.

1. Introduction

The literature has extensively examined the capacity to discern between normal and abnormal occurrences in video sequences. This task involves identifying and separating motion patterns that differ from those acquired during training at inference time. Another inherent difficulty with anomaly identification is the lack of labeled data [1,2]. Since anomalies are, by definition, rare events, they often need to be identified without the availability of labeled samples.

In our scenario, we focus on detecting anomalies from a finer-grained perspective, that is, identifying deviations in regular actions at the level of individual human body joints. The ultimate goal is to uncover subtle behavioral traits that are distinctive to a given motion.

Similar to image anomaly detection, video anomaly detection typically extracts anomaly-free motion features from the recorded routine activities following an unsupervised learning paradigm. Previous efforts [3,4] apply anomaly detection directly on the unprocessed video data. Raw video data may contain unrelated contexts, such as backdrop or illumination changes, which have a detrimental effect on detection accuracy [5]. Therefore, recent solutions [6–8] have switched towards employing skeletal reconstructions to focus directly on human motion.

Many techniques for detecting anomalies in human behavior rely on low-level information and don't comprehend actions semantically [9]. Anomaly detection models frequently exhibit an inadequate level of behavior understanding by using spurious cues and correlations in place of significant, semantic characteristics [10].

Considering the inadequacies of many conventional methods, we present our human action anomaly detection model that leverages unsupervised learning to identify anomalous 3D motion patterns.

To detect the anomalies, we suggest a normalizing flow (NF) [11] based approach for learning high-quality motion representations frame by frame in a sequence. Due to the possibility of near resemblances in the recorded 3D human motion, the motion data may have jitter that makes it difficult to notice subtle changes in action. Although there are numerous joints in the human body, only a small subset of them show noticeable and relevant changes while performing daily activities (such as running, walking, or waving). The remaining joints typically move in a more dependent or restricted manner. Therefore, we considered the frame-by-frame sequence of the motion and determined the likelihood score of the wrong action in each frame. A hybrid anomaly detection framework is proposed with masked auto-encoding and normalizing

☆ This article is part of a Special issue entitled: 'IMAGE_Immersive and Networked Media' published in Signal Processing: Image Communication.

* Corresponding author at: DISI, University of Trento, 38123, Trento, Italy.
E-mail addresses: mahamatissa.choueb@unitn.it (M.I. Choueb), pk.sekharamantry@unitn.it (P.K. Sekharamantry), giulia.martinelli-2@unitn.it (G. Martinelli), francesco.denatale@unitn.it (F. De Natale), nicola.conci@unitn.it (N. Conci).

https://doi.org/10.1016/j.image.2025.117469
Received 16 October 2025; Received in revised form 4 December 2025; Accepted 22 December 2025
Available online 24 December 2025

flows(NFs). Our method allows for systematic scoring of how likely a test motion is under the learned distribution by acquiring the spatiotemporal structure of human motion from normal-activity datasets and modeling the density of latent features. At inference time, we identify subtle anomalous sequences by computing a composite anomaly score that combines latent probability and reconstruction error. NFs are ideal for anomaly scoring because they offer precise probability estimation through invertible transformations: low likelihoods are labeled as out-of-distribution [12].

Practical applications in [13,14] Virtual Reality (VR) and Augmented Reality (AR) environments are considered as essential to better reflect the relevance to immersive media. In these contexts, producing realistic virtual avatars and facilitating organic, intuitive interactions within virtual worlds depend on precise human motion modeling. In addition to boosting user presence and supporting use cases such as virtual training, rehabilitation, and cooperative immersive scenarios, accurate anomaly estimate aids in the maintenance of realistic behavior. These aspects emphasize the importance of the proposed technique within immersive media contexts.

In terms of implementation, our initial step is to train a masked autoencoder (MAE) for skeleton motion sequences. We take advantage of large human motion datasets, such as Mixamo, which offer rich and varied human motion information. After the MAE has been trained, we feed the encoder's embeddings into a NF model. Only latent vectors obtained from normal motion sequences are used to train the NF. The log-likelihood of a test motion's [15] latent embedding can be used to score it during inference. As an output, the NF provides a rigorous probabilistic measure of normality, while the MAE encodes spatiotemporal structure [16]. This dual perspective improves the generalization capacity of anomaly detection systems for human motion and increases sensitivity to minute deviations. Furthermore, this technique performs well with latent vectors generated by motion-focused architectures such as generalized [17] masked autoencoder (MAE) encoders or MOMA [18].

The primary objectives of the proposed work are:

- Design a hybrid anomaly detection system that combines NFs and masked autoencoders to estimate the density of latent representations using just normal motion datasets and simultaneously capture spatiotemporal motion relationships.
- Establish a composite anomaly scoring approach that combines log-likelihood values from NFs with reconstruction error from masked auto-encoding to identify minor deviations that would not be visible using either approach alone.
- Validate the framework at inference time, using a variety of abnormal motions generated by Momask [19], Bamm [20], and Meshcapade [21], to show its versatility, resilience, and enhanced performance over current anomaly detection techniques.

The paper is organised as follows: Section 2 reviews the relevant research on the topic. Section 3 describes the dataset used in the study. Section 4 provides detailed information on the suggested methodology. Section 5 summarizes the experimental setup and findings. Section 6 addresses the findings and their implications. Finally, Section 7 summarizes the paper and discusses future research directions.

2. Related work

Detecting anomalies in motion data, such as skeleton or joint trajectories [22], is a relevant problem in many application areas, including robotics [23], surveillance [24], and animation [25].

Typical techniques involve discovering motion regularity for video segmentation and recognition or rebuilding motion patterns to identify deviations. However, accurate anomaly detection still faces significant challenges in identifying ambiguous behaviors, when abnormal patterns and normal activities [26] differ in subtle details. In contrast to traditional video-based methods, skeleton-based representations use joint locations and temporal dynamics to model human activities, providing a portable, comprehensible, and modality-robust substitute for RGB or optical flow characteristics [18]. Here, a tensor is used to encode each skeleton sequence. This organised representation enables anomaly detection algorithms focus on structural kinematics as, for instance, anomalies in joint trajectories, asymmetries, abnormal limb movements, all of which could be difficult to identify in the raw pixel space [27,28].

Recent approaches integrate the spatiotemporal feature learning and probabilistic modeling to overcome the differences between normal and close variations of abnormal behaviors. A prominent framework for skeleton representation learning has been addressed using transformer architectures [29]. Inter-joint interdependence and long-range temporal dynamics are effectively captured by the self-attention mechanism. The probability of a joint trajectory being within the learned distribution of normal motion can be ascertained using probabilistic models such as normalizing flow [30]. By incorporating reconstruction-driven self-supervision techniques like joint position prediction and masking, it is possible to create strong anomaly detection pipelines that take advantage of contextual reconstruction mistakes as well as statistical probability estimation [31]. Skeleton-based anomaly detection has thus evolved from manually constructed spatiotemporal descriptors to deep, self-supervised transformer-flow hybrids that can capture intricate, joint-level motion patterns.

To this aim, masked modeling has proven particularly effective; here, parts of the input (e.g., motion segments or joints) are intentionally hidden and the network is trained to reconstruct them, thereby learning robust contextual representations without explicit anomaly labels [32]. Conditional Normalizing Flows (CNFs) are efficient generative models that are suitable for structured output learning because they can represent complex distributions with high dimensionality and significant inter-dimensional correlations [33]. One of the recent works in this area is MotionFlow [34], a method that determines the output distributions in an autoregressive manner on the spatiotemporal input features. To develop a probabilistic neural generative method that can simulate the variability observed in high-dimensional structured spatio-temporal data, it blends deterministic and stochastic representations with CNFs. SimMIM [33] is a simple framework for masked image modeling; it consists of an image-based technique that masks picture patches by substituting them with random token vectors of the same dimension. The correlation between tokens is sufficiently strong in the video domain, where data is both spatially and temporally aligned [35–37], and the network can reconstruct the input using most masked frames. In contrast to complete frame masking, joint-level masking keeps much of the skeleton intact at each time step and compels the model to take advantage of both temporal continuity and spatial dependencies. A similar work [38] projected a Spatial-Temporal Masking, which is both at the joint and frame levels. The encoder generates generalizable skeleton characteristics with temporal and spatial relationships. In Masked Action Recognition (MAR) [39], the authors eliminate superfluous processing by only working on a subset of the videos and discarding a portion of the patches. The two key elements in MAR are the masking technique, which hides portions of the input motion sequence, and the reconstruction process, which learns to recover the missing information to improve action understanding [40].

In [41], 2D skeletons with missing joints are used as input to estimate more realistic 3D human positions. The introduction of occlusion guidance, which offers additional details regarding the existence or absence of a joint, addresses missing joints. Additionally, temporal information has been used to more accurately estimate the missing joints. A similar work [42] suggests using synthetic occlusion augmentation coupled with the utilization of spatiotemporal characteristics to address occlusion during training. The authors use an occlusion augmentation to train a spatiotemporal 3D HPE (human pose estimation) model based

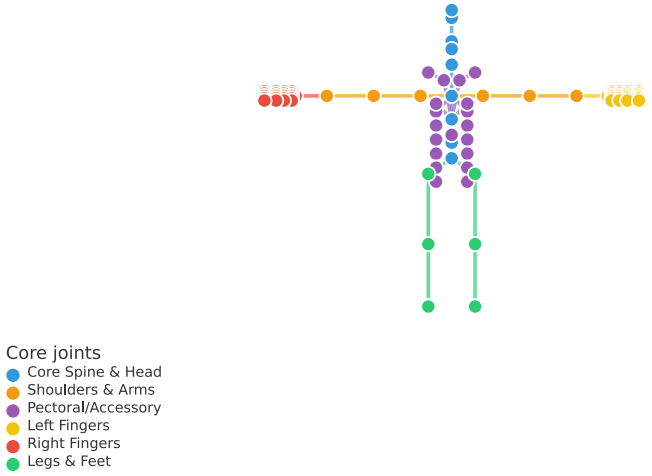


Fig. 1. Representation of Mixamo skeleton showing all (86) joints used in training. The figure highlights the core combined categories, each represented in different color.

on graph convolution and transformers. Comparable works [43,44] also provide generative models of deep features using Gaussian ellipsoidal kernels that generate 3D pose-dependent feature vectors based on a volumetric human representation.

3. Dataset

For the validation of NFlowAD we rely on 04 datasets: Mixamo [45], MoMask [19], Bamm [20], and Meshcapade [21]. More in detail, the Mixamo motion dataset, offers a wide range of typical motions executed by numerous characters. It is used as a source of training data for the anomaly detection framework. Mixamo sequences are highly curated, animation-driven, and cover a broad spectrum of everyday activities, sports, and leisure pursuits. To ensure that learned representations accurately reflect the statistical regularities of natural human movement, training is limited to Mixamo, exposing the model only to motions deemed bio-mechanically valid and consistent. This decision enables the formulation of the next detection step as an out-of-distribution problem, in which variations from these learned distributions are regarded as possible anomalies.

For training, the sequences are 250 frames long sampled to 30 fps, re-targeted to a canonical number of skeleton joints($J = 86$), in line with the Mixamo format, and as illustrated in Fig. 1, and normalized with respect to mean and variance across the dataset, in order to standardize the input space. In order to provide enough variability without adding artificial artifacts, the sequences used for training are chosen to guarantee coverage of a variety of movement patterns, transitions, and upper-body movements. For evaluation, we took into account three additional sources, not seen at training time. The first one is the Meshcapade collection, which includes synthetic human movements; such movements are similar to the ones presented in Mixamo, and include crawling, dancing, squatting, and walking. The sequences provide a useful test of the model's ability to generalize to unseen normal motions, as they are realistic and bio-mechanically plausible. However, they differ from Mixamo in terms of style and kinematic smoothness.

As far as Bamm and MoMask is concerned, those datasets were used to test the framework's capacity to detect unusual motions. Samples from BAMM consist of abnormal behaviors, generated via text prompts, which introduce asymmetric crawling, irregular jumping patterns, and improperly performed squats. Designed to capture human-like deviations from expected motion regularity, these sequences enable the model to be tested in situations that closely resemble incorrect movement patterns encountered in the real world.

MoMask includes motions produced by a programmable human motion synthesis model based on diffusion. MoMask animations are generated from text prompts or semantic action labels. While many sequences appear natural, the generating process can yield minor to pronounced kinematic variations. As such, MoMask is a great source of model-agnostic anomalies: errors result from generative model flaws rather than human execution errors (as in BAMM). Evaluating on MoMask therefore examines the framework's ability to detect algorithmically created abnormalities.

Pre-processing procedures are conducted to ensure uniform representation across all datasets [46]. Every motion is normalized and aligned and adapted regarding its joint skeleton topology. The Mixamo and Meshcapade sequences were used as normal reference data for threshold calibration (one in training and the other in testing), and anomaly detection metrics were calculated using BAMM and MoMask. Instead of learning dataset-specific traits, this sub-division guarantees that the evaluation accurately captures the system's ability to identify out-of-distribution behaviors.

The resulting dataset collection provides a balanced experimental design in which each source performs a unique role. Mixamo offers the key priors that establish the typical motion distribution during training. Meshcapade assesses the model's capacity to generalize to previously encountered, biomechanically valid motions. BAMM assesses robustness against human-performed aberrant or improperly executed actions, capturing realistic deviations from correct movement. MoMask extends this by evaluating the model's sensitivity to generative-model flaws such kinematic inconsistencies and implausible transitions. Together, these datasets provide a controlled and thorough environment for testing normal-motion learning, generalization, and anomaly detection across several types of deviation.

4. Methodology

The proposed method combines masked autoencoding with CNFs to mimic both reconstruction fidelity and probabilistic regularity of human motion. The architecture in Fig. 2 has two stages: (i) a transformer-based masked autoencoder learns robust spatio-temporal representations by reconstructing missing joints and frames, and (ii) a conditional NF head assigns likelihoods to latent embeddings, allowing for anomaly scoring beyond reconstruction errors. This integration enables the detection of anomalies at various granularities, including individual joints to entire sequences.

4.1. Problem formulation

The task of detecting motion anomalies can be described as follows. Given an input sequence $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, where each frame $\mathbf{x}_t \in \mathbb{R}^{J\times 3}$ represents the 3D coordinates of J joints at frame t. The full sequence can be represented as $\mathbf{x} \in \mathbb{R}^{T\times J\times 3}$. Our goal is to determine whether the sequence or any of its components (frames, joints) deviates from the distribution of normal motions.

In our implementation, we assume to have access to a set of *normal* motion sequences and no anomalous samples during the training phase.

4.2. Architecture overview

Our model integrates two complementary modules: (1) a masked autoencoder for robust feature learning, and (2) a CNF head for probabilistic modeling.

- **Masked Autoencoder (MAE):** the encoder–decoder learns to reconstruct missing joints or frames, promoting strong spatial–temporal representations. The masking method forces the network to record relationships between joints and throughout time.

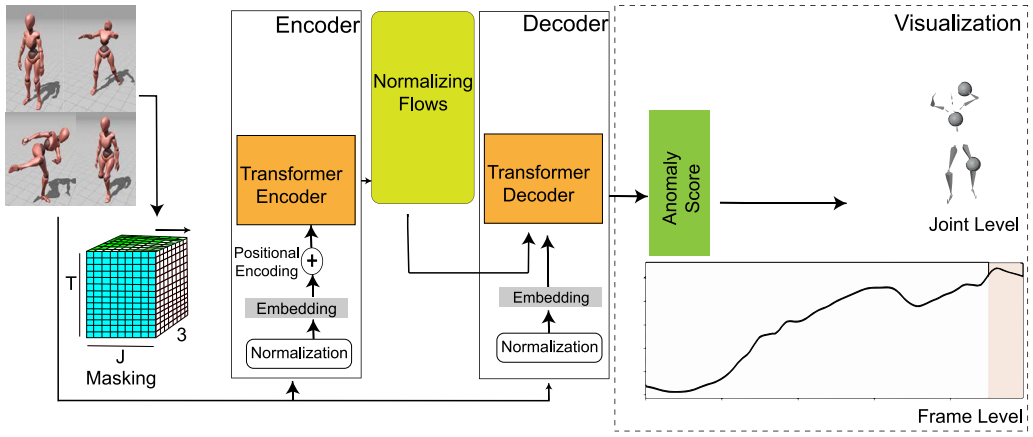


Fig. 2. Architecture of the proposed model. Motion sequences are masked and encoded with a Transformer, then reconstructed using a decoder and NFs, and scored by combining reconstruction and likelihood. Anomalies are identified when scores exceed a threshold.

We use a masked autoencoder technique. During training, a random section of joints ($\rho = 0.3$) are masked, and each frame is linearly embedded as follows:

$$\mathbf{z}_t = W_{\text{emb}}\,\text{vec}(\mathbf{x}_t) + \mathbf{b}_{\text{emb}}, \qquad \mathbf{z}_t \in \mathbb{R}^d, \tag{1}$$

with embedding dimension $d = 512$. A Transformer encoder models spatio-temporal dependencies as:

$$\mathbf{h}_{1:T} = \text{TransformerEnc}(\mathbf{z}_{1:T}), \qquad \mathbf{h}_t \in \mathbb{R}^{d_{\text{model}}}. \tag{2}$$

This encourages the model to capture the interdependence of different elements of the skeleton. As a result, the network learns not only motion dynamics but also the skeleton's underlying kinematics, both of which are required for accurate anomaly identification.

- **Conditional Normalizing Flow (Glow):** The latent embeddings are modeled using a flow-based generative model to estimate the probability density of reconstructed motion. This allows for anomaly scoring not just by reconstruction error, but also by determining if a latent representation is in a high- or low-probability section of the learnt distribution. Let $\mathbf{c}$ represent the pooled encoder context vector obtained from the hidden sequence $\mathbf{h}_{1:T}$. The B invertible transformations $\{f_\theta^{(1)}, \ldots, f_\theta^{(B)}\}$ compose the conditional normalizing flow f_θ:

$$\mathbf{z_t} = f_\theta(\mathbf{h}_t\,;\,\mathbf{c}), \tag{3}$$

with the conditional log-likelihood becoming:

$$\log p_\theta(\mathbf{h}_t \mid \mathbf{c}) = \log p_0(\mathbf{z}_t) + \sum_{b=1}^{B} \log \left| \det \frac{\partial f_\theta^{(b)}}{\partial \mathbf{z}_t^{(b-1)}} \right|. \tag{4}$$

The optimization is carried out by minimizing the negative log-likelihood, defined as $\mathcal{L}_{\text{nll}} = -\log p_\theta(\mathbf{h}_t \mid \mathbf{c})$.

4.3. Inference and anomaly scoring

During inference, anomalies are detected by combining reconstruction error and latent likelihood into a unified score:

$$S(x') = \alpha \cdot \text{MSE}(x', \hat{x}') \; - \; \beta \cdot \log p(\mathbf{z}), \tag{5}$$

where:

- $\text{MSE}(x', \hat{x}')$. represents the per-joint and per-frame reconstruction error between the observed motion x' and the reconstructed motion $\hat{x}'$;
- $p(z')$ represents the likelihood of the latent embedding z' under the flow model;
- $\alpha = 0.6$ and $\beta = 0.4$ are weighting parameters that balance reconstruction fidelity and distributional plausibility. These values are set to favor the reconstruction error over the flow based on validation experiments. The method is not very sensitive to the slight change of α.

This formulation penalizes both deviations in spatial–temporal structure (via reconstruction error) and off-manifold behaviors (via flow likelihood). As a result, anomalies are captured, whether they manifest as locally corrupted poses, abnormal joint trajectories, or globally inconsistent motion dynamics.

4.4. Granularity of anomaly detection

We analyze anomalies at four levels (joint, frame, sequence, clip). Each score is compared against a threshold estimated from validation statistics (e.g., a percentile of the score distribution computed on normal motions).

- **Joint-level.** A joint is marked as anomalous at frame t if its joint score $S_{j,t}$ exceeds the joint threshold τ_{joint}:

  $S_{j,t} > \tau_{\text{joint}}$.

- **Frame-level.** A frame is considered anomalous if its frame score exceeds the frame threshold:

  $S_t > \tau_{\text{frame}}$.

- **Sequence-level.** A sequence is considered anomalous if its global sequence score exceeds the sequence threshold:

  $S_{\text{seq}} > \tau_{\text{seq}}$.

- **Clip-level rule.** Let T be the number of frames in the clip. We compute the fraction of frames that are above the frame anomaly threshold:

  $$S_{\text{frac}} = \frac{\#\{t \in \{1, \ldots, T\} \,:\, S_t > \tau_{\text{frame}}\}}{T}.$$

  A clip is classified as anomalous if this fraction exceeds a minimum value τ_{frac} (e.g., 10%).

5. Experiments

5.1. Implementation details

The input sequences are normalized to zero mean and unit variance across the joints within the frame. Sequence length is fixed to 250 frames during training. The encoder is structured on the transformer

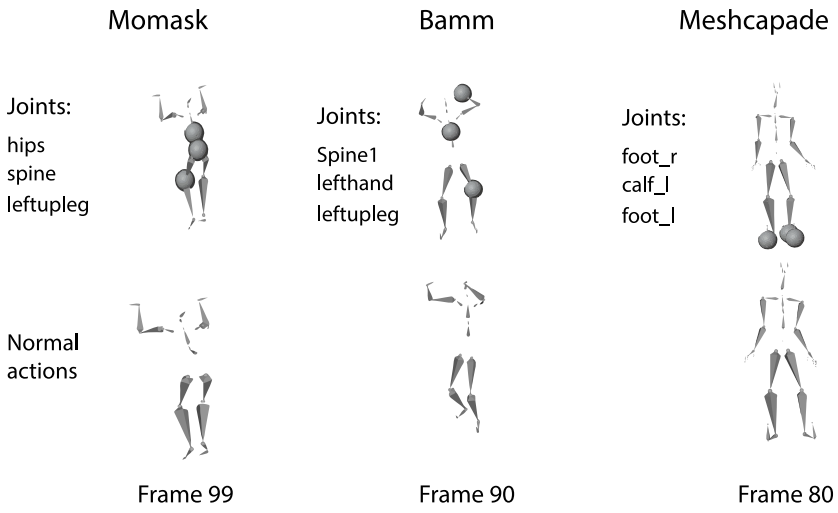


Fig. 3. Joint level detection of top *3* joints per frame with high likelihood score. A subject is performing the braced hang hop up action.

Table 1
Meshcapade clips are treated as normal; BAMM/MoMask clips are treated as anomalous. Values refer to the frame scores s_t likelihood.

Dataset	#Clips	Anomalous clips %	$median(s_t)$	mean(s_t)
Meshcapade (normal)	*30*	*33.33%*	*0.120937*	*0.072924*
BAMM (anomalous)	*30*	*83.33%*	*0.444062*	*0.268886*
MoMask (anomalous)	*30*	*86.67%*	*0.523125*	*0.258414*

encoder with 6 layers, an embedding size of 512, and 8 attention heads. Training is performed for 130 epochs using the Adam optimizer with a learning rate of 1×10^{-4} with an ExponentialLR scheduler. During testing, anomaly thresholds are set using the 97th percentile of reconstruction error and flow likelihood on the validation set, and sequences are fixed respecting each method's sample size.

5.2. Quantitative results

As discussed in the previous sections, our experimental evaluation is performed in a multi-level fashion.

Joint-level localization highlights the precise body parts responsible for anomalies, e.g., misplaced knees in abnormal walking or over-rotated arms during lifting as illustrated in Fig. 3.

Sequence-level evaluation shows robustness in distinguishing abnormal actions from normal motions. Frame-level achieved good results, demonstrating the benefit of combining reconstruction and probabilistic as shown in Table 1 and further demonstrated as a confusion matrix in Fig. 4. The elevated false positive rate observed in Meshcapade likely stems from the dataset's increasing diversity of characters and exaggerated action styles, which introduce distributional shifts not fully represented during training. The model performance in terms of precision and recall is further elaborated in Fig. 5.

As illustrated in Fig. 4, NFs severely penalize deviations from the defined normal manifold. Unusual local positions, abrupt transitions, and stylistic variations are instances of rare but genuine motions that can result in low likelihoods and drive frame scores above our high-percentile criteria.

To reduce missed anomalies, we set a threshold at the 97th percentile of normal validation clips. This design purposely shifts the point of operation towards higher recall/lower false negative rate, which naturally raises false positives.

Across datasets, the score distributions reinforce the intended separation between normal and anomalous motion sources as illustrated in Table 2 for the most challenging clips.

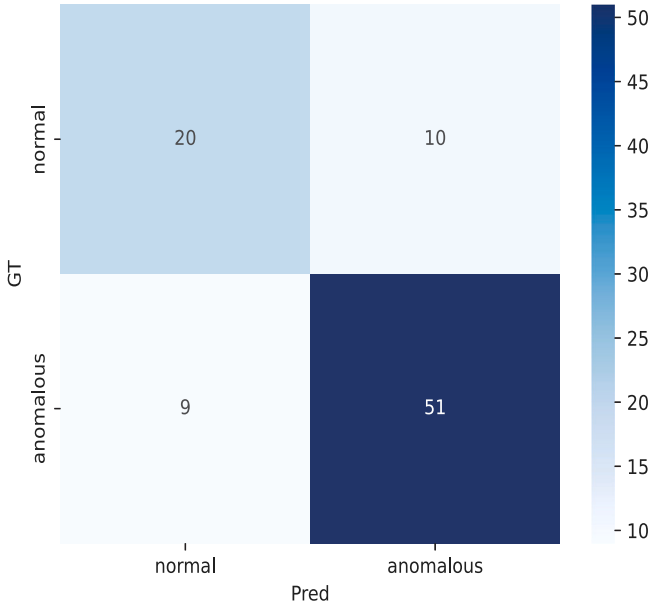


Fig. 4. Confusion matrix obtained from the experimental validation between animations as ground truth (Meshcapade) and samples from Momask and BAMM as anomalous.

MoMask and BAMM dominate the global top hardest and challenging clips in the evaluation test. Several MoMask clips reach $\max_t s_t \in [0.64, 1.39]$; interestingly, a few of the largest $\max_t s_t$ entries (1.236 and 1.388) have $S_{\text{frac}} = 0\%$ less likely anomalous, indicating short, sharp deviations (peaky frames) that the thresholding rule treats as isolated spikes rather than sustained anomalies. In contrast, MoMask entries with $\max_t s_t \approx 0.72, 0.75$ show significant $S_{\text{frac}} \in [33\%, 100\%]$, indicating temporally persistent variances.

BAMM clips also show large maxima ($\max_t s_t \approx 1.074, 1.034$) and, in many cases, with S_{frac} up to 100%. This pattern signals anomalies that are broadly spread across the sequence rather than confined to a handful of frames. This is in line with the expectations, as the animations from BAMM have been curated to explicitly represent abnormal motions.

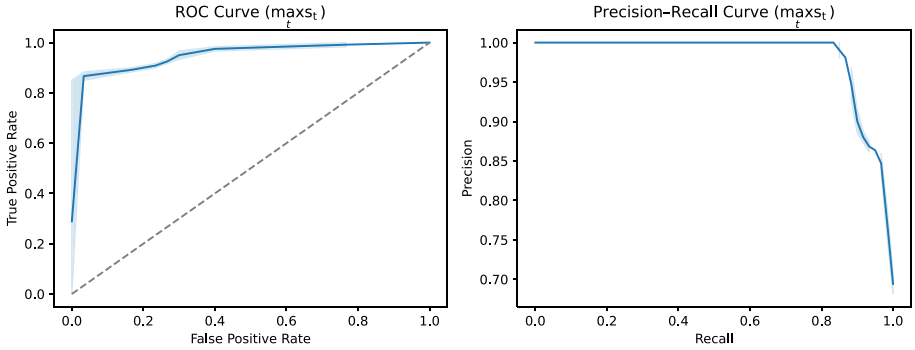


Fig. 5. ROC and Precision-Recall curves were constructed with the frame-level score $\max_t s_t$ as the anomaly indicator. The plots depict the trade-off between true positive and false positive rates at different thresholds (ROC), as well as the balance of precision and recall (PR). These curves supplement the basic quantitative metrics by showing how the detector performs at different decision thresholds, resolving class imbalance and providing a more comprehensive evaluation of anomaly detection ability.

Table 2
Global Top hardest clips across all datasets (sorted by $\max_t s_t$). Anomaly detection results using the fixed frame threshold $\tau_{\text{frame}} = 0.386$ at 97th percentile and clip rule $\mathbf{1}\{\max_t s_t > \tau_{\text{frame}} \vee \text{mean}_t s_t > \tau_{\text{seq}} \vee S_{\text{frac}} \geq 10\%\}$. We consider a clip challenging if $\{\max_t s_t > \tau_{\text{frame}} \vee S_{\text{frac}} \geq 10\%\}$.

Dataset	Clip	$\max_t s_t$	$S_{\text{frac}}\%$	Detected
MoMask	Box Turn with anomaly	1.388	0.0%	N
MoMask	Freehang Drop	1.236	0.0%	N
BAMM	Cross Jumps Rotation	1.074	100.0%	Y
MoMask	Body Jab Cross	1.066	25.0%	Y
BAMM	Drunk Idle	1.034	100.0%	Y
MoMask	Dancing Running	0.749	33.0%	Y
MoMask	Climbing Up Wall	0.721	100.0%	Y
MoMask	Drunk Run Forward	0.642	4.0%	Y
BAMM	Falling Loop	0.605	32.0%	Y
BAMM	End Plank	0.604	41.0%	Y
Meshcapade	Falling Loop	0.291	0.0%	N

Table 3
BAMM (normalized): Anomalous frames and top joint contributors for *braced hang hop up*. All values normalized to $[0, 1]$ using the same range as the temporal heatmap (LeftUpLeg=LUL, LeftHand=LH, RightUpLeg=RUL, RightHand=RH, Spine1=S1, RightFoot=RF, RightShoulder=RS, Hips=H).

Frame	s_t (norm)	Top-1	Top-2	Top-3
90	0.207	LUL: 0.071	LH: 0.019	S1: 0.000
91	0.759	LUL: 0.273	S1: 0.214	LH: 0.144
92	1.000	LUL: 0.286	S1: 0.252	H: 0.207
93	0.919	LH: 0.158	H: 0.155	LUL: 0.140
94	0.676	RUL: 0.222	RH: 0.191	LH: 0.105
95	0.523	RH: 0.463	RUL: 0.365	RF: 0.358
96	0.402	RH: 0.712	RF: 0.617	RUL: 0.474
97	0.269	RH: 0.882	RF: 0.787	RS: 0.570
98	0.139	RH: 0.972	RF: 0.865	RS: 0.652
99	0.000	RH: 0.974	RF: 0.855	RS: 0.661

Table 4
MoMask (normalized): Anomalous frames and top joint contributors for *braced hang hop up*. All values normalized to $[0, 1]$ using the same min/max as the temporal heatmap (Hips=H, LeftUpLeg=LUL, Spine=S).

Frame	s_t (norm)	Top-1	Top-2	Top-3
90	0.000	H: 0.648	S: 0.644	LUL: 0.627
91	0.099	H: 0.651	S: 0.646	LUL: 0.632
92	0.195	H: 0.658	S: 0.651	LUL: 0.640
93	0.289	H: 0.667	S: 0.661	LUL: 0.649
94	0.392	H: 0.676	S: 0.673	LUL: 0.660
95	0.506	H: 0.684	S: 0.684	LUL: 0.671
96	0.610	H: 0.693	S: 0.695	LUL: 0.681
97	0.699	H: 0.699	S: 0.703	LUL: 0.688
98	0.866	H: 0.708	S: 0.712	LUL: 0.698
99	1.000	H: 0.720	S: 0.722	LUL: 0.709

5.3. Qualitative results

To complement the obtained quantitative metrics, we also provide some visualizations, which could be more intuitive for the reader, in Fig. 6.

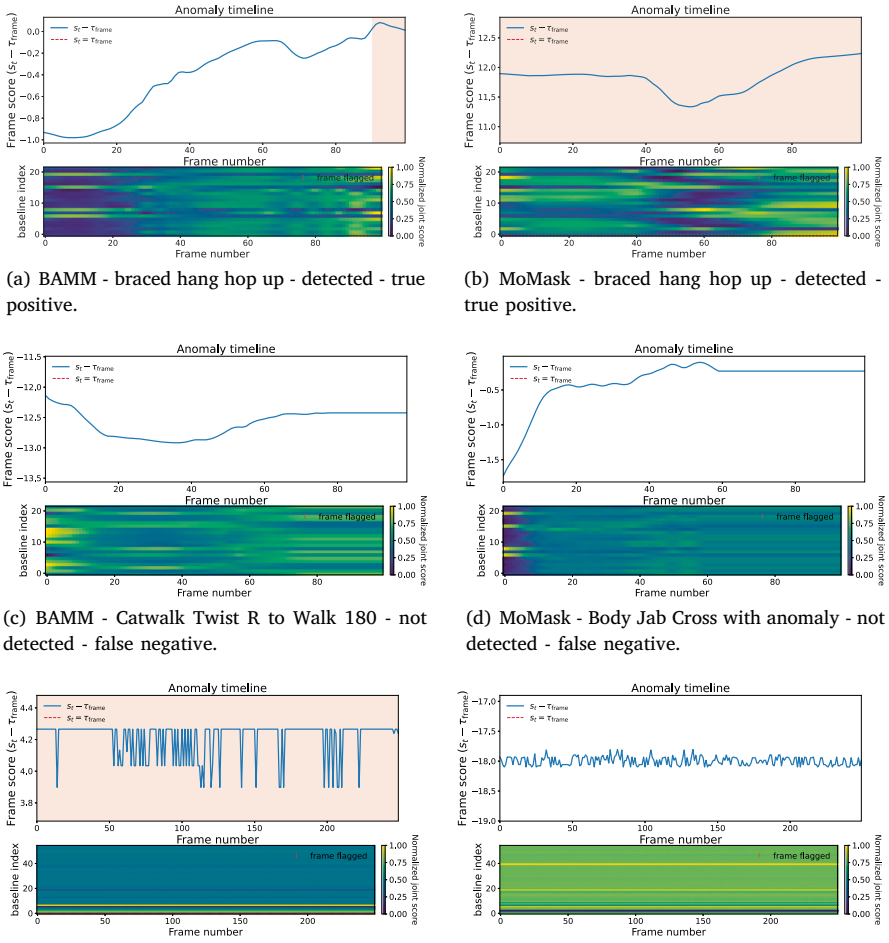
The reported diagrams show the evolution of the joint's behavior over time. The joints exhibiting anomalies are highlighted, with severity proportional to the anomaly score. This is further explained in Table 3, and Table 4 as in subfigures (a) and (b) for illustrated in the temporal-heatmap. These aforementioned provide top-k joints that present high likelihood anomalies contributing to the abnormal video sequence at clip level.

Further, in Fig. 6 Frame-level anomaly bars and overlays present temporal dynamics of anomalies, enabling intuitive interpretation of abnormal sequences. For instance, in corrupted motion sequences, the joints are most likely to produce distorted movements, which are identified as anomalies, indicating implausible or unnatural poses. In each sub-figure, the top part shows the frame-level anomaly score over time, with the red dashed line indicating the threshold estimated globally for each sequence. The cross point of the curve beyond the threshold highlights an anomalous frame. In the bottom part, the temporal heatmap emphasizes the joint scores per frame. We have on the x-axis the number of frames and the corresponding index of the baseline method. A color map is provided to indicate the intensity of the anomaly for these joints. A high intensity comparable to 1 is likely anomalous, and a low intensity depicted with dark color is normal. An anomalous frame in this sequence is flagged on this temporal heatmap following the axis.

6. Discussion

The findings in this study show that our encoder–decoder method offers a strong framework for motion anomaly detection in conjunction with quantile-based thresholding. The system intentionally treats normality with caution. Although external animated characters from Meshcapade were not visible during training, they consistently produced less false positives. This supports our hypothesis that learning a generalizable manifold of plausible human movement is actually possible.

The fact that false positives vary widely within the tested samples is not surprising, as they are diverse in terms of the number of characters with varied actions in the dataset, as mentioned earlier. Even though they are semantically valid, some high-velocity or stylistically exaggerated actions cause frame-level anomalies. The model may misinterpret valid activities at high speeds or exaggerations as abnormalities. This demonstrates a trade-off: when it is too sensitive, it detects expressive



(a) BAMM - braced hang hop up - detected - true positive.

(b) MoMask - braced hang hop up - detected - true positive.

(c) BAMM - Catwalk Twist R to Walk 180 - not detected - false negative.

(d) MoMask - Body Jab Cross with anomaly - not detected - false negative.

(e) Meshcapade - braced hang hop up (normal) - detected - false positive.

(f) Meshcapade - Body Jab Cross (normal) - not detected - true negative.

Fig. 6. Signaling of anomalies for different video sequences. Each sub-figure includes a plot showing the anomaly scores over time and a joint-level visualization where highlighted lines indicate anomaly intensity at the joint level. Each horizontal line refers to a tracked joint over frames.

motions; when it is too tolerant, it may miss actual anomalies. This balance could improve with more adaptive thresholding or conditioning based on action type.

The results for the anomalous data sources (BAMM, MoMask) confirm a significant separation from the learnt normal distribution. Both datasets contain motions that have been intentionally curated to exhibit physically implausible or semantically corrupted behaviors, and our method identifies them with near-perfect recall at the sequence level. MoMask distinguishes between isolated peaks and persistent anomalies, with $\max_t s_t$ values occasionally above the threshold but no lasting deviations. This demonstrates the effectiveness of the temporal persistence rule ($S_{\text{frac}} \geq 10\%$), which prevents over-flagging of transient frame anomalies. The ability to see the joints that lead to an anomaly is key for interpretability, and this is further reinforced by the reconstruction-based visualization.

Though, there are limitations. First, the use of a global quantile threshold relates performance to the statistical distribution of the validation set; as a result, cross-dataset deployment may benefit from calibration methods such as extreme value theory or Bayesian uncertainty modeling. Second, while it may be tempting to introduce a taxonomy of anomaly types (e.g., biomechanical mistakes, semantic inconsistencies, or temporal jitter), we understand that doing so could turn the problem into a classification task. Our current binary approach remains appropriate since anomaly identification is basically about spotting anomalies from predicted behaviors, even without labels. What is important is robustly detecting deviations; fine-grained categorization would require additional supervision and may dilute the generalization benefits of an unsupervised anomaly detector.

7. Conclusion

This research proposed a reconstruction-driven anomaly detection framework for human motion sequences that captures both temporal dynamics and spatial structure by combining a masked autoencoder and a flow-based decoder. After being trained only on normal Mixamo motions, the system achieves high recall on curated anomalous samples (BAMM, MoMask) and generalizes well to unseen normal data as in Meshcapade. When combined with temporal persistence rules, frame-level scores provide accurate clip-level anomaly determinations. Crucially, the ability to visualize per-joint reconstruction errors improves interpretability, enabling practitioners to identify the structural causes of abnormality.

Three main conclusions are supported by the empirical evaluation: (i) the model learns a reasonable variety of plausible human motions, which leads to a low false-positive rate on external normal data; (ii) abnormal datasets are successfully separated, with anomaly scores significantly above threshold levels; and (iii) reconstruction-based visualization offers an interpretable connection between numerical anomaly scores and their kinematic appearances.

In future work, we aim to extend the framework in different directions. First, adaptive thresholding strategies conditioned on motion

type or velocity could reduce false positives on stylistically extreme but valid motions. Second, integrating semantic priors (e.g., action labels or biomechanical constraints) could enhance anomaly categorization beyond binary classification.

CRediT authorship contribution statement

Mahamat Issa Choueb: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Conceptualization. **Praveen Kumar Sekharamantry:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **Giulia Martinelli:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **Francesco De Natale:** Supervision, Investigation, Funding acquisition. **Nicola Conci:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Nicola Conci reports financial support was provided by European Union under the Italian National Recovery and Resilience Plan (NRRP). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART") Mission 4 Component 2 - CUP C69J24000180004.

Data availability

Data will be made available on request.

References

[1] Y. Cao, X. Xu, J. Zhang, Y. Cheng, X. Huang, G. Pang, W. Shen, A survey on visual anomaly detection: Challenge, approach, and prospect, 2024, arXiv preprint arXiv:2401.16402, Online: https://arxiv.org/abs/2401.16402.

[2] Z. Ye, Y. Chen, H. Zheng, Understanding the effect of bias in deep anomaly detection, 2021, arXiv preprint arXiv:2105.07346, Online: https://arxiv.org/abs/2105.07346.

[3] W. Li, V. Mahadevan, N. Vasconcelos, Anomaly detection and localization in crowded scenes, IEEE Trans. Pattern Anal. Mach. Intell. 36 (1) (2014) http://dx.doi.org/10.1109/TPAMI.2013.111.

[4] F. Jiang, J. Yuan, S.A. Tsaftaris, A.K. Katsaggelos, Anomalous video event detection using spatiotemporal context, Comput. Vis. Image Underst. 115 (3) (2011) http://dx.doi.org/10.1016/j.cviu.2010.10.008.

[5] M. Sabokrou, M. Fayyaz, M. Fathy, R. Klette, Deep-cascade: Cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes, IEEE Trans. Image Process. 26 (4) (2017) http://dx.doi.org/10.1109/TIP.2017.2670780.

[6] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, S. Venkatesh, Learning regularity in skeleton trajectories for anomaly detection in videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, http://dx.doi.org/10.1109/CVPR.2019.01227.

[7] R. Rodrigues, N. Bhargava, R. Velmurugan, S. Chaudhuri, Multi-timescale trajectory prediction for abnormal human activity detection, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, WACV, 2020, http://dx.doi.org/10.1109/WACV45572.2020.9093633.

[8] A. Flaborea, L. Collorone, G.M. D'Amely Di Melendugno, S. D'Arrigo, B. Prenkaj, F. Galasso, Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, Oct., 2023, pp. 10284–10295, http://dx.doi.org/10.1109/ICCV51070.2023.00947.

[9] H. Mu, R. Sun, G. Yuan, Y. Wang, Abnormal human behavior detection in videos: A review, Inf. Technol. Control. 50 (3) (2021) 522–545.

[10] J. Kauffmann, L. Ruff, G. Montavon, K.-R. Müller, The clever hans effect in anomaly detection, 2020, arXiv preprint arXiv:2006.10609, Online: https://arxiv.org/abs/2006.10609.

[11] A. Zanfir, E.G. Bazavan, H. Xu, W.T. Freeman, R. Sukthankar, C. Sminchisescu, Weakly supervised 3D human pose and shape reconstruction with normalizing flows, in: Lecture Notes in Computer Science, 2020, http://dx.doi.org/10.1007/978-3-030-58539-6_28.

[12] Z. Hu, et al., MSAttnFlow: Normalizing flow for unsupervised anomaly detection with multi-scale attention, Pattern Recognit. 161 (2025) 111220, http://dx.doi.org/10.1016/j.patcog.2024.111220.

[13] Z. Zhang, Y. Wang, B. Wu, S. Chen, Z. Zhang, S. Huang, W. Zhang, M. Fang, L. Chen, Y. Zhao, Motion avatar: Generate human and animal avatars with arbitrary motion, 2024, arXiv preprint arXiv:2405.11286, Online: https://arxiv.org/abs/2405.11286.

[14] E. Makled, C. Gerhardt, T. Schwandt, F. Weidner, W. Broll, Evaluating behavioral realism in AR and VR: A comparison of single-point IK and full-body motion capture virtual humans, 2025, http://dx.doi.org/10.21203/rs.3.rs-6354819/v1.

[15] A. Bousse, et al., Maximum-likelihood joint image reconstruction and motion estimation with misaligned attenuation in TOF-PET/CT, Phys. Med. Biol. 61 (3) (2016) http://dx.doi.org/10.1088/0031-9155/61/3/L11.

[16] C.J. Reed, et al., Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning, in: Proceedings of the IEEE International Conference on Computer Vision, 2023, http://dx.doi.org/10.1109/ICCV51070.2023.00378.

[17] H. Yan, Y. Liu, Y. Wei, Z. Li, G. Li, L. Lin, SkeletonMAE: Graph-based masked autoencoder for skeleton sequence pre-training, in: Proceedings of the IEEE International Conference on Computer Vision, 2023, http://dx.doi.org/10.1109/ICCV51070.2023.00516.

[18] G. Martinelli, N. Garau, N. Bisagno, N. Conci, Skeleton-aware motion retargeting using masked pose modeling, 2025, pp. 287–303, http://dx.doi.org/10.1007/978-3-031-92387-6_21.

[19] C. Guo, Y. Mu, M.G. Javed, S. Wang, L. Cheng, Momask: Generative masked modeling of 3D human motions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1900–1910.

[20] E. Pinyoanuntapong, M.U. Saleem, P. Wang, M. Lee, S. Das, C. Chen, BAMM: Bidirectional autoregressive motion model, in: European Conference on Computer Vision, Springer, 2024, pp. 172–190.

[21] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M.J. Black, SMPL: A skinned multi-person linear model, in: *Seminal Graphics Papers: Pushing the Boundaries, vol. 2*, 2023, pp. 851–866.

[22] Z. Qin, et al., Fusing higher-order features in graph neural networks for skeleton-based action recognition, IEEE Trans. Neural Netw. Learn. Syst. 35 (4) (2024) http://dx.doi.org/10.1109/TNNLS.2022.3201518.

[23] Q. Yang, F. Xu, J. Leng, Abnormal actions detection of robotic arm via 3D convolution neural network and support vector data description, J. Intell. Fuzzy Syst. 42 (6) (2022) http://dx.doi.org/10.3233/JIFS-212468.

[24] H. Yuan, J.H. Lee, S. Zhang, Research on simulation of 3D human animation vision technology based on an enhanced machine learning algorithm, Neural Comput. Appl. 35 (6) (2023) http://dx.doi.org/10.1007/s00521-022-07083-x.

[25] J. Arunnehru, G. Chamundeeswari, S.P. Bharathi, Human action recognition using 3D convolutional neural networks with 3D motion cuboids in surveillance videos, in: Procedia Computer Science, 2018, http://dx.doi.org/10.1016/j.procs.2018.07.059.

[26] E.M.C.L. Ekanayake, Y. Lei, C. Li, Crowd density level estimation and anomaly detection using multicolumn multistage bilinear convolution attention network (MCMS-BCNN-Attention), Appl. Sci. (Switzerland) 13 (1) (2023) http://dx.doi.org/10.3390/app13010248.

[27] Y. Mao, J. Deng, W. Zhou, Y. Fang, W. Ouyang, H. Li, Masked motion predictors are strong 3D action representation learners, in: Proceedings of the IEEE International Conference on Computer Vision, 2023, http://dx.doi.org/10.1109/ICCV51070.2023.00934.

[28] J. Zhang, L. Lin, J. Liu, Prompted contrast with masked motion modeling: Towards versatile 3D action representation learning, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, http://dx.doi.org/10.1145/3581783.3611774.

[29] A. Snoun, T. Bouchrika, O. Jemai, View-invariant 3D skeleton-based human activity recognition based on transformer and spatio-temporal features, in: International Conference on Pattern Recognition Applications and Methods, 2022, http://dx.doi.org/10.5220/0010895300003122.

[30] Z. You, et al., A unified model for multi-class anomaly detection, Adv. Neural Inf. Process. Syst. (2022).

[31] Y. Zhou, X. Xu, J. Song, F. Shen, H.T. Shen, Msflow: Multiscale flow-based framework for unsupervised anomaly detection, IEEE Trans. Neural Netw. Learn. Syst. 36 (2) (2025) http://dx.doi.org/10.1109/TNNLS.2023.3344118.

[32] Q. Li, X. Huang, et al., MM 2023 - Proceedings of the 31st ACM International Conference on Multimedia, 2023, http://dx.doi.org/10.1145/3581783.3612496.

[33] Y. Lu, B. Huang, Structured output learning with conditional generative flows, in: Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020, http://dx.doi.org/10.1609/aaai.v34i04.5940.

[34] M. Zand, A. Etemad, M. Greenspan, Flow-based spatio-temporal structured prediction of motion dynamics, IEEE Trans. Pattern Anal. Mach. Intell. 45 (11) (2023) http://dx.doi.org/10.1109/TPAMI.2023.3296446.

[35] Z. Tong, Y. Song, J. Wang, L. Wang, VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training, Adv. Neural Inf. Process. Syst. (2022).

[36] C. Feichtenhofer, H. Fan, Y. Li, K. He, Masked autoencoders as spatiotemporal learners, Adv. Neural Inf. Process. Syst. (2022).

[37] L. Wang, et al., VideoMAE V2: Scaling video masked autoencoders with dual masking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2023, http://dx.doi.org/10.1109/CVPR52729.2023.01398.

[38] W. Wu, Y. Hua, C. Zheng, S. Wu, C. Chen, A. Lu, Skeletonmae: Spatial-temporal masked autoencoders for self-supervised skeleton action recognition, in: Proceedings of the IEEE International Conference on Multimedia and Expo Workshops, ICMEW, 2023, http://dx.doi.org/10.1109/ICMEW59549.2023.00045.

[39] Z. Qing, et al., MAR: Masked autoencoders for efficient action recognition, IEEE Trans. Multimed. 26 (2024) http://dx.doi.org/10.1109/TMM.2023.3263288.

[40] Z. Zhou, X. Liu, Masked autoencoders in computer vision: A comprehensive survey, IEEE Access 11 (2023) 113560–113579, http://dx.doi.org/10.1109/ACCESS.2023.3323383.

[41] M. Ghafoor, A. Mahmood, Quantification of occlusion handling capability of a 3D human pose estimation framework, IEEE Trans. Multimed. 25 (2023) http://dx.doi.org/10.1109/TMM.2022.3158068.

[42] S. Banik, P. Gschoßmann, A.M. García, A. Knoll, Occlusion robust 3D human pose estimation with StridedPoseGraphFormer and data augmentation, in: Proceedings of the International Joint Conference on Neural Networks, 2023, http://dx.doi.org/10.1109/IJCNN54540.2023.10191355.

[43] Y. Zhang, P. Ji, A. Wang, J. Mei, A. Kortylewski, A. Yuille, 3D-aware neural body fitting for occlusion robust 3D human pose estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2023, pp. 9365–9376, http://dx.doi.org/10.1109/ICCV51070.2023.00862.

[44] B. Biggs, S. Ehrhardt, H. Joo, B. Graham, A. Vedaldi, D. Novotny, 3D multi-bodies: Fitting sets of plausible 3D human models to ambiguous image data, Adv. Neural Inf. Process. Syst. (2020).

[45] Z. Guo, J. Xiang, K. Ma, W. Zhou, H. Li, R. Zhang, Make-it-animatable: An efficient framework for authoring animation-ready 3d characters, 2024, arXiv preprint arXiv:2411.18197.

[46] S.B. Kotsiantis, D. Kanellopoulos, P.E. Pintelas, Data preprocessing for supervised learning, Int. J. Comput. Sci. 1 (2) (2006) 111–117.