Skeletal Video Anomaly Detection Using Deep Learning: Survey, Challenges, and Future Directions

Pratik K. Mishra ⓘ, Alex Mihailidis ⓘ, and Shehroz S. Khan ⓘ

Abstract—The existing methods for video anomaly detection mostly utilize videos containing identifiable facial and appearance-based features. The use of videos with identifiable faces raises privacy concerns, especially when used in a hospital or community-based setting. Appearance-based features can also be sensitive to pixel-based noise, straining the anomaly detection methods to model the changes in the background and making it difficult to focus on the actions of humans in the foreground. Structural information in the form of skeletons describing the human motion in the videos is privacy-protecting and can overcome some of the problems posed by appearance-based features. In this paper, we present a survey of privacy-protecting deep learning anomaly detection methods using skeletons extracted from videos. We present a novel taxonomy of algorithms based on the various learning approaches. We conclude that skeleton-based approaches for anomaly detection can be a plausible privacy-protecting alternative for video anomaly detection. Lastly, we identify major open research questions and provide guidelines to address them.

Index Terms—Skeleton, body joint, human pose, anomaly detection, video.

I. INTRODUCTION

ANOMALOUS events pertain to unusual or abnormal actions, behaviours or situations that can lead to health, safety and economical risks [1]. Anomalous events, by definition, are largely unseen and not much is known about them in advance [2]. Due to their rarity, diversity and infrequency, collecting labeled data for anomalous events can be very difficult or costly [1], [3]. With the lack of predetermined classes and a few labelled data for anomalous events, it can be very hard to train supervised machine learning models [1]. Therefore, a general approach in majority of anomaly detection algorithms is to train a model that can best represent the 'normal' events or actions, and any deviations from it can be flagged as an unseen anomaly [4]. Anomalous behaviours among humans can be attributed at an individual level (e.g., falls [5]) or multiple people in a scene (e.g., pedestrian crossing [6], violence in a crowded mall [7]). In the context of video-based anomaly detection, the general approach is to train a model to learn the patterns of actions or behaviours of individual(s), background and other semantic information in the normal activities videos, and identify significant deviations in the test videos as anomalies. However, anomaly detection is a challenging task due to the lack of labels and often times the unclear definition of an anomaly [2].

The majority of video-based anomaly detection approaches use RGB videos where the people in the scene are identifiable. While using RGB camera-based systems in public places (e.g., malls, airports) is generally acceptable, the situation can be very different in personal dwelling, community, residential or clinical settings [8]. In a home or residential setting (e.g., nursing homes), individuals or patients can be monitored in their personal space that may breach their privacy. The lack of measures to deal with the privacy of individuals can be a bottleneck in the adoption and deployment of the anomaly detection-based systems [9]. However, monitoring of people with physical, cognitive or aging issues is also important to improve their quality of life and care. Therefore, as a trade-off, privacy-protecting video modalities can fill that gap and be used in these settings to save lives and improve patient care. Wearable devices face compliance issues among certain populations, where people may forget or in some cases refuse to wear them [10]. Some of the privacy-protecting camera modalities that has been used in the past for anomaly detection involving humans include depth cameras [5], [11], thermal cameras [12], and infrared cameras [13], [14]. While these modalities can partially or fully obfuscate an individual's identity, they require specialized hardware or cameras and can be expensive to be used by general population. Skeletons extracted from RGB camera streams using pose estimation algorithms provide a suitable solution of privacy protection over RGB and other types of cameras [15]. Skeleton tracking only focuses on body joints and ignores facial identity, full body scan or background information. The pixel-based features in RGB videos that mask important information about the scene are sensitive to noise resulting from illumination, viewing direction and background clutter, resulting in false positives when detecting anomalies [16]. Furthermore, due to redundant information present in these features (e.g., background), there is an increased burden on methods to model the change in those areas of the scene rather than focus on the actions of humans in the foreground. Extracting information specific to human actions can not only provide a privacy-protecting solution, but can also help to filter out the background-related noise in the videos and aid the model to focus on key information for detecting abnormal events related to human behaviour. The skeletons

represent an efficient way to model the human body joint positions over time and are robust to the complex background, illumination changes, and dynamic camera scenes [17]. In addition to being privacy-protecting, skeleton features are compact, well-structured, semantically rich, and highly descriptive about human actions and motion [17]. Anomaly detection using skeleton tracking is an emerging area of research as awareness around privacy of individuals and their data grows. However, skeleton-based approaches may not be sufficient for situations that explicitly need facial information for analysis, including emotion recognition [18], [19], pain detection [20] or remote heart monitoring [21], to name a few.

In recent years, deep learning methods have been developed to use skeletons for different applications, such as action recognition [48], medical diagnosis [24], and sports analytics [49]. The use of skeletons for anomaly detection in videos is an underexplored area, and concerted research is needed [24]. The human skeletons can help in developing privacy-preserving solutions for private dwellings, crowded/public areas, medical settings, rehabilitation centers and long-term care homes to detect anomalous events that impact health and safety of individuals. Use of this type of approach could improve the adoption of video-based monitoring systems in homes and residential settings. However, there is a paucity of literature on understanding the existing techniques that use skeleton-based anomaly detection approaches. We identify this gap in the literature and present one of the first surveys on the recent advancements in using skeletons for anomaly detection in videos. We identified the major themes in existing work and present a novel taxonomy that is based on how these methods learn to detect anomalous events. We also discuss the applications where these approaches were used to understand their potential in bringing these algorithms in a personal dwelling, or long-term care scenario.

II. LITERATURE SURVEY

We adopted a narrative literature review for this work. The following keywords (and their combinations) were used to search for relevant papers – skeleton, human pose, body pose, body joint, anomaly detection, and video. These keywords were searched on scholarly databases, including Google Scholar, IEEE Xplore, Elsevier and Springer. We mostly reviewed papers between year 2016 to 2023; therefore, the list may not be comprehensive. In this review, we only focus on the recent deep learning-based algorithms for skeletal video anomaly detection and do not include traditional machine learning-based approaches. There are works [50], [51] on detecting anomalous behaviour using supervised approaches, however, it is outside the scope of this review as it focuses on unsupervised anomaly detection approaches. We did not adopt the systematic or scoping review search protocol for this work; therefore, our literature review may not be exhaustive. However, we tried our best to include the latest development in the field to be able to summarize their potential and identify challenges. In this section, we provide a survey of skeletal deep learning video anomaly detection methods. We present a novel taxonomy to study the skeletal video anomaly approaches based on learning approaches

into four broad categories, i.e., reconstruction, prediction, their combinations and other specific approaches. Table I provides a summary of 29 relevant papers, based on the taxonomy, found in our literature search. Unless otherwise specified, the values in the last column of the table refer to AUC(ROC) values corresponding to each dataset in the reviewed paper. Six papers use reconstruction approach, six papers use prediction approach, seven papers use a combination of reconstruction and prediction approaches, five papers use a combination of reconstruction and clustering approaches, and five papers use other specific approaches.

A. Reconstruction Approaches

In the reconstruction approaches, generally, an autoencoder (AE) or its variant model is trained on the skeleton information of only normal human activities. During training, the model learns to reconstruct the samples representing normal activities with low reconstruction error. Hence, when the model encounters an anomalous sample at test time, it is expected to give high reconstruction error.

Gatt et al. [22] used Long Short-Term Memory (LSTM) and 1-Dimensional Convolution (1DConv)-based AE models to detect abnormal human activities, including, but not limited to falls, using skeletons estimated from videos of a publicly available dataset. Temuroglu et al. [23] proposed a skeleton trajectory representation that handled occlusions and an AE framework for pedestrian abnormal behaviour detection. The pedestrian video dataset used in this work was collected by the authors, where the training dataset was composed of normal walking, and the test dataset was composed of normal and drunk walking. The pose skeletons were treated to handle occlusions using the proposed representation and combined into a sequence to train an AE. They compared the results of occlusion-aware skeleton keypoints input with keypoints without occlusion flags, keypoint image heatmaps and raw pedestrian image inputs. The authors used average of recall and specificity to evaluate the models due to the unbalanced dataset and found that occlusion-aware input achieved the highest results. Suzuki et al. [24] trained a Convolutional AE (CAE) on good gross motor movements in children and detected poor limb motion as an anomaly. Motion time-series images [52] were obtained from skeletons estimated from the videos of kindergarten children participants. The motion time-series images were fed as input to a CAE, which was trained on only the normal data. The difference between the input and reconstructed pixels was used to localize the poor body movements in anomalous frames. Jiang et al. [25] presented a message passing Gated Recurrent Unit (GRU) encoder-decoder network to detect and localize the anomalous pedestrian behaviours in videos captured at the grade crossing. The field-collected dataset consisted of over 50 hours of video recordings at two selected grade crossings with different camera angles. The skeletons were estimated and decomposed into global and local components before being fed as input to the encoder-decoder network. The localization of the anomalous pedestrians within a frame was done by identifying the skeletons with reconstruction error higher than the empirical threshold.

TABLE I
SUMMARY OF REVIEWED PAPERS

Learning approach	Paper	Datasets used	Experimental Setting	Number of people in scene	Type of anomalies	Pose estimation algorithm	Model input	Model type	Anomaly score	Eval. metric AUC(ROC) (or other)
Reconstruction	Gatt et al. [22]	UTD-MHAD	Indoor	Single	Irregular body postures	Openpose, Posenet	Skeleton keypoints	1DConv-AE, LSTM-AE	Reconstruction error	AUC(PR)=0.91, F score=0.98
	Temuroglu et al. [23]	Custom	Outdoor	Multiple	Drunk walking	Openpose	Skeleton keypoints	AE	Reconstruction error	Average of recall and specificity=0.91
	Suzuki et al. [24]	Custom	—	Single	Poor body movements in children	Openpose	Motion time-series images	CAE	Reconstruction error	Accuracy=99.3, F score=0.99
	Jiang et al. [25]	Custom	Outdoor	Multiple	Abnormal pedestrian behaviours at grade crossings	Alphapose	Skeleton keypoints	GRU Encoder-Decoder	Reconstruction error	0.82
	Song et al. [26]	Custom	Outdoor	Multiple	Abnormal pedestrian behaviours at grade crossings	Openpose	Skeleton keypoints	GAN	Discriminator score	0.89
	Fan et al. [27]	CUHK Avenue, UMN	Indoor and Outdoor	Multiple	Anomalous human behaviours	Alphapose	Video frame, Skeleton keypoints	GAN	Reconstruction error of video frame	0.88 0.99
Prediction	Rodrigues et al. [28]	IITB-Corridor, ShanghaiTech, CUHK Avenue	Outdoor	Multiple	Abnormal human activities	Openpose	Skeleton keypoints	Multi-timescale 1DConv encoder-decoder	Prediction error from different timescales	0.67 0.76 0.83
	Luo et al. [16]	ShanghaiTech, CUHK Avenue	Outdoor	Multiple	Irregular body postures	Alphapose	Skeleton joints graph	Spatio-Temporal GCN	Prediction error	0.74 0.87
	Zeng et al. [29]	UCSD Pedestrian, ShanghaiTech, CUHK Avenue, IITB-Corridor	Outdoor	Multiple	Anomalous human behaviours	HRNet	Skeleton joints graph	Hierarchical Spatio-Temporal GCN	Weighted sum of prediction errors from different levels	0.98 0.82 0.87 0.7
	Fan et al. [30]	ShanghaiTech, CUHK Avenue	Outdoor	Multiple	Anomalous human actions	Alphapose	Skeleton keypoints	GRU feed forward network	Prediction error	0.83 0.92
	Pang et al. [31]	ShanghaiTech, CUHK Avenue	Outdoor	Multiple	Anomalous human actions	Alphapose	Skeleton keypoints	Transformer	Prediction error	0.77 0.87
	Huang et al. [32]	ShanghaiTech, CUHK Avenue, IITB-Corridor	Outdoor	Multiple	Anomalous human behaviours	HRNet	Skeleton joints graph	Spatio-Temporal Graph Transformer	Max of prediction errors of local & global graphs	0.83 0.89 0.77
Reconstruction+ Prediction	Morais et al. [17]	ShanghaiTech, CUHK Avenue	Outdoor	Multiple	Anomalous human actions	Alphapose	Skeleton keypoints	GRU Encoder-Decoder	Weighted sum of reconstruction and prediction errors	0.73 0.86
	Boekhoudt et al. [7]	ShanghaiTech, HR Crime	Indoor and Outdoor	Multiple	Human and Crime related anomalies	Alphapose	Skeleton keypoints	GRU Encoder-Decoder	Weighted sum of reconstruction and prediction errors	0.73 0.6
	Li and Zhang [33]	ShanghaiTech	Outdoor	Multiple	Abnormal pedestrian behaviours	Alphapose	Skeleton keypoints	GRU Encoder-Decoder	Weighted sum of reconstruction and prediction errors	0.75
	Li et al. [34]	ShanghaiTech, CUHK Avenue	Outdoor	Multiple	Human-related anomalous events	Alphapose	Skeleton joints graph	GCAE with embedded LSTM	Sum of max reconstruction and prediction errors	0.76, EER=30.7 0.84, EER=20.7
	Wu et al. [35]	ShanghaiTech, CUHK Avenue	Outdoor	Multiple	Abnormal human actions	Alphapose	Skeleton joints graph, Confidence scores	GCN	Confidence score weighted sum of reconstruction, prediction and SVDD errors	0.77 0.85
	Luo et al. [36]	ShanghaiTech, IITB-Corridor	Outdoor	Multiple	Human-related video anomalies	—	Skeleton joints graph	Memory Enhanced Spatial-Temporal GCAE	Sum of max reconstruction and prediction errors	0.78 0.69
	Li et al. [37]	ShanghaiTech, CUHK Avenue	Outdoor	Multiple	Human-related anomalous events	Alphapose	Skeleton keypoints	Memory-augmented GAN	Sum of max reconstruction and prediction errors	0.75, EER=31.7 0.84, EER=22.6
Reconstruction+ Clustering	Markovitz et al. [38]	ShanghaiTech, NTU-RGB+D, Kinetics-250	Indoor and Outdoor	Multiple	Anomalous human actions	Alphapose, Openpose	Skeleton joints graph	GCAE, Deep clustering	Dirichlet process mixture model score	0.75 0.85 0.74
	Cui et al. [39]	ShanghaiTech	Outdoor	Multiple	Human pose anomalies	—	Skeleton joints graph	GCAE, Deep clustering	Dirichlet process mixture model score	0.77
	Liu et al. [40]	ShanghaiTech, CUHK Avenue	Outdoor	Multiple	Anomalous human behaviours	Alphapose	Skeleton joints graph	GCAE, Deep clustering	Dirichlet process mixture model score	0.79 0.88
	Chen et al. [41]	ShanghaiTech, CUHK Avenue	Outdoor	Multiple	Anomalous human behaviours	Alphapose	Skeleton joints graph	Multiscale spatial temporal attention GCN	Skeleton cluster anomaly score	0.76, EER=31.1 0.88, EER=19.2
	Yan et al. [42]	ShanghaiTech, UCF-Crime, NTU-RGB+D	Outdoor	Multiple	Anomalous human actions	Openpose	Skeleton joints graph	GCAE, Deep clustering	Skeleton cluster anomaly score	0.77 0.76 0.77
Clustering	Yang et al. [43]	UCSD Pedestrian 2, ShanghaiTech	Outdoor	Multiple	Anomalous human behaviours and objects	Alphapose	Skeleton joints graph, Numerical features	GCN	Skeleton cluster + Object anomaly score	0.93 0.82
	Javed et al. [44]	ShanghaiTech, UCF-Crime, NTU-RGB+D	Outdoor	Multiple	Anomalous human actions	—	Video frame, Skeleton joints graph	GCN, Deep clustering	Dirichlet process mixture model score	0.81 0.86 0.88
Iterative self-training	Nanjun et al. [45]	ShanghaiTech, CUHK Avenue	Outdoor	Multiple	Human-related anomalous events	Alphapose	Skeleton joints graph, Numerical features	GCN	Self-trained fully connected layers output	0.72, EER=34.1 0.82, EER=23.9
Multivariate gaussian distribution	Tani and Shibata [46]	ShanghaiTech	Outdoor	Multiple	Anomalous human behaviours	Openpose	Skeleton joints graph	GCN, Multivariate gaussian distribution	Mahalanobis distance	0.77
Prompt-guided zero-shot learning	Sato et al. [47]	RWF-2000 Kinetics-250	Outdoor	Multiple	Abnormal-human behavior events	PPN HRNet	Skeleton keypoints	Residual multilayer perceptron	Joint probability score	Accuracy=90.3 0.79

They manually removed wrongly detected false skeletons as they claim that the wrong detection issue was observed at only one grade crossing. However, an approach of manual removal of false skeletons is impractical in many real world applications where the data is very large, making the need of an automated false skeleton identification and removal step imperative. In their following work [26], the authors improved the performance of detecting abnormal pedestrian behaviors at grade crossings using a generative adversarial network (GAN)-based framework. Two LSTM-based branches within the generator were used to analyze both local and global motion patterns simultaneously, reconstructing the corresponding inputs in the temporal domain.

The discriminator was a fully connected neural network and produced a score representing the likelihood of inputs being an anomaly. Fan et al. [27] proposed an anomaly detection framework which consisted of two pairs of generator and discriminator. The generators were trained to reconstruct the normal video frames and the corresponding skeletons, respectively. The discriminators were trained to distinguish the original and reconstructed video frames and the original and reconstructed skeletons, respectively. The video frames and corresponding extracted skeletons served as input to the framework during training; however, at test time, decision was made based on only reconstruction error of video frames.

Challenges: AEs or their variants are widely used in many video-based anomaly detection methods [5]. The choice of the right architecture to model the skeletons is very important. Further, being trained on the normal data, they are expected to produce higher reconstruction error for the abnormal inputs than the normal inputs, which has been adopted as a criterion for identifying anomalies. However, this assumption does not always hold in practice, that is, the AEs can generalize well that it can also reconstruct anomalies well, leading to false negatives [53].

B. Prediction Approaches

In prediction approaches, a network is generally trained to learn the normal human behaviour by predicting the skeletons at the next time step(s) using the skeletons representing normal human actions at past time steps. During testing, the test samples with high prediction errors are flagged as anomalies as the network is trained to predict only the skeletons representing normal actions.

Rodrigues et al. [28] suggested that abnormal human activities can take place at different timescales, and the methods that operate at a fixed timescale (frame-based or video-clip-based) are not enough to capture the wide range of anomalies occurring with different time duration. They proposed a multi-timescale 1DConv encoder-decoder network where the intermediate layers were responsible to generate future and past predictions corresponding to different timescales. The network was trained to make predictions on normal activity skeletons input. The prediction errors from all timescales were combined to get an anomaly score to detect abnormal activities. Luo et al. [16] proposed a spatio-temporal Graph Convolutional Network (GCN)-based prediction method for skeleton-based video anomaly detection. The body joints were estimated and built into skeleton graphs, where the body joints formed the nodes of the graph. The spatial edges connected different joints of a skeleton, and temporal edges connected the same joints across time. A fully connected layer was used at the end of the network to predict future skeletons. Zeng et al. [29] proposed a hierarchical spatio-temporal GCN, where high-level representations encoded the trajectories of people and the interactions among multiple identities while low-level skeleton graph representations encoded the local body posture of each person. The method was proposed to detect anomalous human behaviours in both sparse and dense scenes. The inputs were organized into spatio-temporal skeleton graphs

whose nodes were human body joints from multiple frames and fed to the network. The network was trained on the input skeleton graph representations of normal activities. Optical flow fields and size of skeleton bounding boxes were used to determine sparse and dense scenes. For dense scenes with crowds, higher weights were assigned to high-level representations while for sparse scenes, the weights of low-level graph representations were increased. During testing, the prediction errors from different branches were weighted and combined to obtain the final anomaly score. Fan et al. [30] proposed a GRU feed-forward network that was trained to predict the next skeleton using past skeleton sequences and a loss function that incorporated the range and speed of the predicted skeletons. Pang et al. [31] proposed a skeleton transformer to predict future pose components in video frames and considered error between predicted pose components and corresponding expected values as anomaly score. They applied a multi-head self-attention module to capture long-range dependencies between arbitrary pairwise pose components and the temporal convolutional layer to concentrate on local temporal information. Huang et al. [32] proposed a spatio-temporal graph transformer to encode the hierarchical graph embeddings of human skeletons for jointly modeling the interactions between individuals and the correlations among body joints within a single individual. Input to the transformer was provided as global and local graphs. Each node in the global graph encoded the speed of an individual as well as the relative position and interaction relations between individuals. Each local graph encoded the pose of an individual.

Challenges: In these methods, it is difficult to choose how far in future (or past) the prediction should be made to achieve optimum results. This could potentially be determined empirically; however, in the absence of a validation set such solutions remain elusive. The future prediction-based methods can be sensitive to noise in the past data [54]. Any small changes in the past can result in significant variation in prediction, and not all of these changes signify anomalous situations.

C. Combinations of Learning Approaches

In this section, we discuss the existing methods that utilize a combination of different learning approaches, namely, reconstruction and prediction approaches, and reconstruction and clustering approaches.

1) Combination of Reconstruction and Prediction Approaches: Some skeletal video anomaly detection methods utilize a multi-objective loss function consisting of both reconstruction and prediction errors to learn the characteristics of skeletons signifying normal behaviour and identify skeletons with large errors as anomalies. Morais et al. [17] proposed a method to model the normal human movements in surveillance videos using human skeletons and their relative positions in the scene. The human skeletons were decomposed into two sub-components: global body movement and local body posture. The global movement tracked the dynamics of the whole body in the scene, while the local posture described the skeleton configuration. The two components were passed as

input to different branches of a message passing GRU single-encoder-dual-decoder-based network. The branches processed their data separately and interacted via cross-branch message passing at each time step. Each branch had an encoder, a reconstruction-based decoder and a prediction-based decoder. The network was trained using normal data, and during testing, a frame-level anomaly score was generated by aggregating the anomaly scores of all the skeletons in a frame to identify anomalous frames. In order to avoid the inaccuracy caused by incorrect detection of skeletons in video frames, the authors left out video frames where the skeletons cannot be estimated by the pose estimation algorithm. Hence, the results in this work was not a good representation of a real-world scenario, which often consists of complex-scenes with occluding objects and overlapping movement of people. Boekhoudt et al. [7] utilized the network proposed by Morais et al. [17] for detecting human crime-based anomalies in videos using a newly proposed crime-based video surveillance dataset. Similar to the work by Morais et al. [17], Li and Zhang [33] proposed a dual branch single-encoder-dual-decoder GRU network that was trained on normal behaviour skeletons estimated from pedestrian videos. The two decoders were responsible for reconstructing the input skeletons and predicting future skeletons, respectively. However, unlike the work by Morais et al. [17], there was no provision of message passing between the branches. Li et al. [34] proposed a single-encoder-dual-decoder architecture established on a spatio-temporal Graph CAE (GCAE) embedded with a LSTM network in hidden layers. The two decoders were used to reconstruct the input skeleton sequences and predict the unseen future sequences, respectively, from the latent vectors projected via the encoder. The sum of maximum reconstruction and prediction errors among all the skeletons within a frame was used as anomaly score for detecting anomalous frames. Wu et al. [35] proposed a GCN-based encoder-decoder architecture that was trained using normal action skeleton graphs and keypoint confidence scores as input to detect anomalous human actions in surveillance videos. The skeleton graph input was decomposed into global and local components. The network consisted of three encoder-decoder pipelines: the global pipeline, the local pipeline and the confidence score pipeline. The global and local encoder-decoder-based pipelines learned to reconstruct and predict the global and local components, respectively. The confidence score pipeline learned to reconstruct the confidence scores. Further, a Support Vector Data Description (SVDD)-based loss was employed to learn the boundary of the normal action global and local pipeline encoder output in latent feature space. The network was trained using a multi-objective loss function, composed of a weighted sum of skeleton graph reconstruction and prediction losses, confidence score reconstruction loss and multi-center SVDD loss. Luo et al. [36] proposed a single-encoder-dual-decoder memory enhanced spatial-temporal GCAE network, where spatial-temporal graph convolution was used to encode discriminative features of skeleton graphs in spatial and temporal domains. The memory module recorded patterns for normal behaviour skeletons. Further, the encoded representation was not fed directly into the reconstructing and predicting decoders but was used as a query to retrieve the most relevant memory

items. The memory module was used to restrain the reconstruction and prediction capability of the network on anomalies. Li et al. [37] proposed memory-augmented Wasserstein GAN with gradient penalty to predict future human skeleton trajectories from a given past and reconstruct the given past simultaneously. While the discriminator attempted to fit the Wasserstein distance between the distribution of real and generated samples, the generator tried to minimize the Wasserstein distance to draw the distribution of real and generated samples closer. A memory module was applied in the generator to mitigate the strong generalization ability.

2) Combination of Reconstruction and Clustering Approaches: Some skeletal video anomaly detection methods utilize a two-stage approach to identify anomalous human actions using spatio-temporal skeleton graphs. In the first pre-training stage, a GCAE-based model is trained to minimize the reconstruction loss on input skeleton graphs. In the second fine-tuning stage, the latent features generated by the pre-trained GCAE encoder is fed to a clustering layer and a Dirichlet Process Mixture model is used to estimate the distribution of the soft assignment of feature vectors to clusters. Finally at the test time, the Dirichlet normality score is used to identify the anomalous samples. Markovitz et al. [38] identified that anomalous actions can be broadly classified in two categories, fine and coarse-grained anomalies. Fine-grained anomaly detection refers to detecting abnormal variations of an action, e.g., abnormal type of walking. Coarse-grained anomaly detection refers to defining particular normal actions and regarding other actions as abnormal, such as determining dancing as normal and gymnastics as abnormal. They utilized a spatio-temporal GCAE to map the skeleton graphs representing normal actions to a latent space, which was soft assigned to clusters using a deep clustering layer. The soft-assignment representation abstracted the type of data (fine or coarse-grained) from the Dirichlet model. After pre-training of GCAE, the latent feature output of the encoder and clusters were fine-tuned by minimizing a multi-objective loss function consisting of both the reconstruction loss and clustering loss. They leveraged ShanghaiTech [55] dataset to test the performance of their proposed method on fine-grained anomalies, and NTU-RGB+D [56] and Kinetics-250 [57] datasets for coarse-grained anomaly detection performance evaluation. Cui et al. [39] proposed a semi-supervised prototype generation-based method for video anomaly detection to reduce the computational cost associated with graph-embedded networks. Skeleton graphs for normal actions were estimated from the videos and fed as input to a shift spatio-temporal GCAE to generate features. It was not clear which pose estimation algorithm was used to estimate the skeletons from video frames. The generated features were fed to the proposed prototype generation module designed to map the features to prototypes and update them during the training phase. In the pre-training step, the GCAE and prototype generation module were optimized using a loss function composed of reconstruction loss and generation loss of prototypes. In the fine-tuning step, the entire network was fine-tuned using a multi-objective loss function, composed of reconstruction loss, prototype generation loss and cluster loss. Later, Liu et al. [40] used self-attention augmented graph convolutions for detecting

abnormal human behaviours based on skeleton graphs. Skeleton graphs were fed as input to a spatio-temporal self-attention augmented GCAE and latent features were extracted from the encoder part of the trained GCAE. After pre-training of GCAE, the entire network was fine-tuned using a multi-objective loss function consisting of both the reconstruction loss and clustering loss. Chen et al. [41] proposed a multiscale spatial temporal attention GCN, which included an encoder to extract features, a reconstruction decoder branch to optimize encoder, and a clustering layer branch to obtain anomaly scores. During training, the decoder is used to optimize the encoder by minimizing the reconstruction error. However, during testing, the decoder is discarded, and only the clustering layer is used to generate the anomaly score. It used three scales of human skeleton graphs, namely, joint, part and limb. Spatial attention graph convolution operation was carried out on each scale, and the output features of three scales were weighted and summed to constitute the multiscale skeleton features. Yan et al. [42] proposed a deep memory storage clustering method based on GCAE to implement the real-time updating of pseudo-labels and network parameters. It consisted of a feature extraction, autoencoder, clustering, memory storage, self-supervision and scoring modules. The feature extraction module [38] and the autoencoder module were used to form the reconstructed pose sequence. The reconstructed sequence was then sent to the memory storage module for storage, and the soft cluster assignment was performed on each sample through the k-means clustering method [58]. The autoencoder, clustering, and memory storage modules were used to update the pseudo-labels and network parameters iteratively.

Challenges: The combination-based methods can carry the limitations of the individual learning approaches, as described in Section II-A and II-B. Further, in the absence of a validation set, it is difficult to determine the optimum value of combination coefficients in a multi-objective loss function.

D. Other Approaches

This section discusses the methods that leveraged a pre-trained deep learning model to encode latent features from the input skeletons and used approaches such as, clustering and multivariate gaussian distribution, in conjunction for detecting human action-based anomalies in videos.

Yang et al. [43] proposed a two-stream fusion method to detect anomalies pertaining to body movement and object positions. YOLOv3 [59] was used to detect people and objects in the video frames. Subsequently, skeletons were estimated from the video frames and passed as input to a spatio-temporal GCN, followed by a clustering-based fully connected layer to generate anomaly scores for skeletons. The information pertaining to the bounding box coordinates and confidence score of the detected objects was used to generate object anomaly scores. Finally, the skeleton and object normality scores were combined to generate the final anomaly score for a frame. Nanjun et al. [45] used the skeleton features estimated from the videos for pedestrian anomaly detection using an iterative self-training strategy. The training set consisted of unlabelled normal and anomalous video sequences. The

skeletons were decomposed into global and local components, which were fed as input to an unsupervised anomaly detector, iForest [60], to yield the pseudo anomalous and normal skeleton sets. The pseudo sets were used to train an anomaly scoring module, consisting of a spatial GCN and fully connected layers with a single output unit. As part of the self-training strategy, new anomaly scores were generated using previously trained anomaly scoring module to update the membership of skeleton samples in the skeleton sets. The scoring module was then re-trained using updated skeleton sets, until the best scoring model was obtained. However, the paper doesn't discuss the criteria to decide the best scoring model. Tani and Shibata [46] proposed a framework for training a frame-wise Adaptive GCN (AGCN) for action recognition using single frame skeletons and used the features extracted from the AGCN to train an anomaly detection model. As part of the proposed framework, a pretrained action recognition model [61] was used to identify the frames with large temporal attention in the Kinetics-skeleton dataset [62] as the action frames to train the AGCN. Further, the trained AGCN was used to extract features from the normal behaviour skeletons identified in the ShanghaiTech Campus dataset [17] to model a multivariate gaussian distribution. During testing, the Mahalanobis distance was used to calculate the anomaly score under the multivariate gaussian distribution. Sato et al. [47] proposed a user prompt-guided zero-shot learning framework for the detection of abnormal human behaviour events. A multi-layer perceptron feature extractor was pretrained on large-scale action recognition datasets [63], [64] using contrastive learning between the skeleton features and the text embeddings extracted from action class names. The distribution of skeleton features of the normal actions was modeled during training while freezing the weights of feature extractor. During inference, the anomaly score was computed using distribution and the text prompts of an unseen action. Javed et al. [44] proposed a unified framework for learning suitable frames of interest to cut down on redundant data and a two-stream feature block with a hyper-gated fusion model to take advantage of skeleton graph and video frame features. Soft assignments were later processed through a clustering layer, where probabilities were assigned to the instances and a normality score was calculated using the Dirichlet Process Mixture model [65].

Challenges: The performance of these methods rely on the pre-training strategy of the deep learning models used to learn the latent features and the choice of training parameters for the subsequent machine learning models.

III. DISCUSSION

This section leverages Table I and synthesizes the information and trends that can be inferred from the existing work on skeletal video anomaly detection.

A. Datasets

ShanghaiTech [55] and CUHK Avenue [66] were the most frequently used video datasets to evaluate the performance of the skeletal video anomaly detection methods. The ShanghaiTech dataset has videos of people walking along a sidewalk of the ShanghaiTech university and anomalous frames contain bikers,

TABLE II
CHARACTERISTICS OF SKELETAL VIDEO ANOMALY DETECTION DATASETS

Dataset	Total frames	Training frames	Test frames	Anomalous events	Camera views	Available annotations	Anomalies
CUHK Avenue [66]	30652	15328	15324	47	1	Temporal, Pixel-wise, Track-ID	Throwing object, child skipping, wrong direction, bag on grass
IITB-Corridor [28]	483566	301999	181567	108278 frames	1	Temporal	Protest, unattended baggage, biker, fighting, chasing, loitering, suspicious object, hiding, playing with ball
ShanghaiTech [55]	317398	274515	42883	130	13	Temporal, Pixel-wise	Throwing object, jumping, pushing, bikers, loitering, climbing
UCF-Crime [67]	1900 videos	1610 videos	290 videos	950 videos	—	Temporal	Abuse, arrest, arson, assault, accident, burglary, explosion, fighting, robbery, shooting, stealing, shoplifting, vandalism
UCSD Pedestrian [6]	18560	9350	9210	77	2	Temporal, Pixel-wise, Track-ID	Biker, skater, cart, wheelchair, walk across walkways
UMN [68]	3855	—	—	11	—	Temporal	Abandoned or thrown objects, unusual crowd activity, camera tampering

Fig. 1. One normal and one anomalous frame from each of the skeletal video anomaly detection datasets.

skateboarders and people fighting. It has 330 training videos and 107 test videos. However, not all the anomalous activities are related to humans. A subset of the ShanghaiTech dataset that contained anomalous activities only related to humans was termed as HR ShanghaiTech and was used in many papers. The CUHK Avenue dataset consists of short video clips looking at the side of a building with pedestrians walking by it. Concrete columns that are part of the building cause some occlusion. The dataset contains 16 training videos and 21 testing videos. The anomalous events comprise of actions such as "throwing papers', "throwing bag', "child skipping', "wrong direction' and "bag on grass'. Similarly, a subset of the CUHK Avenue dataset containing anomalous activities only related to humans, called

HR Avenue, has been used to evaluate the methods. Other video datasets that have been used include UTD-MHAD [69], UMN [68], UCSD Pedestrian [6], IITB-Corridor [28], UCF-Crime [67], HR Crime [7], NTU-RGB+D [56], RWF-2000 [70] and Kinetics-250 [57]. Table II presents a summary of the characteristics of these datasets and Fig. 1 presents one normal and one anomalous frame from these datasets. Among the datasets used in the reviewed papers, some of the datasets were originally not meant for but instead adopted for the task of video anomaly detection. Hence, we only provide details for the datasets that were originally meant for video anomaly detection in Table II and Fig. 1. From the type of anomalies present in these datasets, it can be inferred that the

existing skeletal video anomaly detection methods have been evaluated mostly on individual human action-based anomalies. Hence, it is not clear how well can they detect anomalies that involve interactions among multiple individuals or interactions among people and objects.

B. Number of People in the Scene

Most of the papers (27 out of 29), detected anomalous human actions for multiple people in the video scene. Other two papers detected irregular body postures and poor body movements in children, respectively, for single person in the video scene. The usual approach was to estimate the skeletons for the people in the scene using a pose estimation algorithm, and calculate anomaly scores for each of the skeletons. The maximum anomaly score among all the skeletons within a frame was used to identify the anomalous frames. A single video frame could contain multiple people, among which not all of them were performing anomalous actions. Hence, taking the maximum anomaly score of all the skeletons helped to nullify the effect of people with normal actions on the final decision for the frame. Further, calculating anomaly scores for individual skeletons helped to localize the source of anomaly within a frame.

C. Fields of Application

The definition of anomalous human behaviours can differ across various applications. While most of the existing papers focused on detecting anomalous human behaviours in general, five papers focused on detecting anomalous behaviours for specific applications, including drunk walking [23], poor body movements in children [24], abnormal pedestrian behaviours at grade crossings [25], [26] and crime-based anomalies [7]. Moreover, the nature of anomalous behaviours can vary depending upon various factors, such as span of time, crowded scenes, and specific action-based anomalies. Some papers identified and addressed the need to detect specific types of anomalies, namely, multi-timescale anomalies occurring over different time duration [28], anomalies in both sparse and crowded scenes [29], fine and coarse-grained anomalies [38] and body movement and object position anomalies [43].

D. Choice of Pose Estimation Algorithm

Alphapose [71] and Openpose [72] were the most common choice of pose estimation algorithm for extraction of skeletons for the people in the scene. Other pose estimation methods that have been used were Posenet [73], PPN [74] and HRNet [75]. However, in general, the papers did not provide any rationale behind their choice of the pose estimation algorithm.

E. Model Type

The type of models used in the papers can broadly be divided into two types, sequence-based and graph-based models. The sequence-based models that have been used include 1DConv-AE, LSTM-AE, GRU, and Transformer. These models treated skeleton keypoints for individual people across multiple frames as time series input. The graph-based models that have been used involve GCAE and GCN. The graph-based models received spatio-temporal skeleton graphs for individual people as input. The spatio-temporal graphs were constructed by considering body joints as the nodes of the graph. The spatial edges connected different joints of a skeleton, and temporal edges connected the same joints across time.

F. Evaluation Metrics

The choice of a suitable threshold for anomaly detection can vary across different applications as most applications come with different costs for false alarms and missed anomalies [76], [77]. As such, having a metric capable of evaluating the performance of anomaly detection methods across diverse application scenarios, or equivalently, across a wide array of decision thresholds is highly desirable. The Area Under Curve (AUC) of Receiver Operating Characteristic (ROC) curve computes the fraction of detected anomalies, averaged over the full range of decision thresholds. It is the standard evaluation measure used in anomaly detection [76] and also the most common metric used to evaluate the performance among the existing skeletal video anomaly detection methods. The highest AUC(ROC) values reported for the commonly used ShanghaiTech [55] and CUHK Avenue [66] datasets across different methods in Table I were 0.83 and 0.92, respectively. A direct comparison may not be possible due to the difference in the experimental setup and train-test splits across the reviewed methods; however, it gives some confidence on the viability of these approaches for skeletal video anomaly detection. Other performance evaluation metrics include F score, accuracy, Equal Error Rate (EER) and AUC of Precision-Recall (PR) Curve. EER signifies the percentage of misclassified frames when the false positive rate equals to the miss rate on the ROC curve. While AUC(ROC) can provide a good estimate of the classifier's performance over different thresholds, it can be misleading in case the data is imbalanced [78]. In anomaly detection scenario, it is common to have imbalance in the test data, as the anomalous behaviours occur infrequently, particularly in many medical applications [79], [80]. The AUC(PR) value provides a good estimate of the classifier's performance on imbalanced datasets [78]; however, only one of the papers used AUC(PR) as an evaluation metric.

IV. CHALLENGES

A. Pose Estimation Algorithms

In general, the efficiency of the skeletal video anomaly detection algorithms depends upon the accuracy of the skeletons estimated by the pose-estimation algorithm. If the pose estimation algorithm misses certain joints or produces artifacts in the scene, then it can increase the number of false alarms. There are various challenges associated with estimating skeletons from video frames [81]: (i) complex body configuration causing self-occlusions and complex poses, (ii) diverse appearance, including clothing, and (iii) complex environment with occlusion from other people in the scene, various viewing angles, distance from camera and truncation of parts in the camera view. This can lead to a poor approximation of skeletons and can negatively impact

Fig. 2. Openpose (top row) and alphapose (bottom row) output on different datasets.

the performance of the anomaly detection algorithms. Further, there is an associated high cost of powerful hardware required for extracting skeletons using deep learning methods. Methods have been proposed to address some of these challenges [82], [83]; however, extracting skeletons in complex environments remains a difficult problem. The two most commonly used pose estimation algorithms in the reviewed papers are Openpose [72] and Alphapose [71]. Multi-person pose estimation can be categorized into top-down and bottom-up methods [81]. Top-down methods [71], [84] usually employ human detectors to obtain bounding boxes for humans in the input frame and then utilize existing single-person pose estimators to predict body joints. This method highly depends upon the precision of human detection algorithms, and the run-time is proportional to the number of persons in the input frame. Bottom-up methods [72] directly approximate all the body joints of all the humans in the input frame and assemble them into individual skeletons. However, the grouping of joints in a complex scene is a challenging task. Openpose is a bottom-up method and Alphapose is a top-down method. Fig. 2 presents the skeleton output of openpose and alphapose on different dataset frames. Some of the existing methods manually remove inaccurate and false skeletons [17], [25] to train the model, which is impractical in many real-world applications where the amount of available data is very large. There is a need for an automated false skeleton identification and removal step when estimating skeletons from videos.

B. Types of Anomalies

The anomalous human behaviours of interest and their difficulty of detection can vary depending upon the definition of anomaly, application, time span of the anomalous actions, and presence of single/multiple people in the scenes. For example, in the case of driver anomaly detection application, the anomalous behaviours can include talking on the phone, dozing off or drinking [14]. The anomalous actions can span over different time lengths, ranging from few seconds to hours or days, e.g., jumping and falls [85] are short-term anomalies, while loitering and social isolation [86] are long-term events. More focus is needed on developing methods that can identify both short and long-term anomalies. Sparse scene anomalies can be described as anomalies in scenes with less number of humans, while dense

scene anomalies can be described as anomalies in crowded scenes with a large number of humans [29]. It is comparatively difficult to identify anomalous behaviours in dense scenes than sparse scenes due to tracking multiple people and finding their individual anomaly scores [17]. Thus, there is a need to develop methods that can effectively identify both sparse and dense scene anomalies. With the development of algorithms for handling different types of anomalies, there is a need for datasets composed of the specific type of anomalies to ensure efficient training and evaluation. This can be handled by either having separate datasets for specific types of anomalies or general datasets with a distribution of multiple types of anomalies.

C. Hardware

The skeletons collected using Microsoft Kinect (depth) camera has been used in the past studies [87], [88]. However, the defunct production of the Microsoft Kinect camera [89] has led to hardware constraints in the further development of skeletal anomaly detection approaches. Other commercial products include Vicon [90] with optical sensors and TheCaptury [91] with multiple cameras. But they function in very constrained environments or require special markers on the human body. New cameras, such as 'Sentinare 2' from AltumView [92], circumvent such hardware requirements by directly processing videos on regular RGB cameras and transmitting skeletons information in real-time.

D. Tracking Skeletons

The existing approaches for skeletal video anomaly detection involve spatio-temporal skeleton graphs [16] or temporal sequences [17], which are constructed by tracking an individual across multiple frames. However, this is challenging in scenarios where there are multiple people within a scene. The entry and exit of people in the scene, overlapping of people during movement and presence of occluding objects make tracking people across frames a very challenging task.

E. Choice of Threshold

There can be deployment issues in these methods because the choice of threshold is not clear. In the absence of any

validation set (containing both normal and unseen anomalies) in an anomaly detection setting, it is very hard to fine-tune an operating threshold using just the training data (comprising of normal activities only). To handle these situations, outliers within the normal activities can be used as a proxy for unseen anomalies [85], [93]; however, inappropriate choices can lead to increased false alarms or missed alarms. Domain expertise can be utilized to adjust a threshold, which may not be available in many cases.

F. Decision Granularity

There is a need to address the challenges associated with the granularity and the decision-making time of the skeletal video anomaly detection methods for real-time applications. The existing methods mostly output decisions on a frame level, which becomes an issue when the input to the method is a real-time continuous video stream at multiple frames per second. This can lead to alarms going off multiple times a second, which can be counter-productive. One solution is for the methods to make decisions on a time-window basis, each window of length of a specified duration. However, this brings in the question about the optimal length of each decision window. A short window is impractical as it can lead to frequent and repetitive alarms, while a long window can lead to missed alarms, and delayed response and intervention. Domain knowledge can be used to make a decision about the length of decision windows.

V. FUTURE DIRECTIONS

Skeletons can be used in conjunction with optical flow [94] to develop privacy-protecting approaches to jointly learn from temporal and structural modalities. Approaches based on federated learning (that do not combine individual data, but only the models) can further improve the privacy of these methods [95]. Segmentation masks [96] can be leveraged in conjunction with skeletons to occlude humans while capturing the information pertaining to scene and human motion to develop privacy-protecting anomaly detection approaches.

The skeletons signify motion and posture information for the individual humans in the video; however, they lack information regarding human-human and human-object interactions. Information pertaining to interaction of the people with each other and the objects in the environment is important for applications such as, violence detection [7], theft detection [7] and agitation detection [80] in care home settings. Skeletons can be used to replace the bodies of the participants, while keeping the background information in video frames [97] to analyze both human-human and human-object interaction anomalies. Further, object bounding boxes can be used in conjunction with human skeletons to model human-object interaction while preserving the privacy of humans in the scene. The information from other modalities (e.g. wearable devices) along with skeleton features can be used to develop multi-modal anomaly detection methods to improve the detection performance. Further, the generated embeddings of relevant supervised approaches [98], [99] can be used to fine tune skeletal video anomaly detection models.

As can be seen in Table I, the existing skeletal video anomaly detection methods and available datasets focus towards detecting irregular body postures [16], and anomalous human actions [31] in mostly outdoor settings, and not in proper healthcare settings, such as personal homes and long-term care homes. This a gap towards real world deployment, as there is a need to extend the scope of detecting anomalous behaviours using skeletons to in-home and care home settings, where privacy is a very important concern. This can be utilized to address important applications, such as fall detection [100], agitation detection [80], [97], and independent assistive living. This will help to develop supportive homes and communities and encourage autonomy and independence among the increasing older population and dementia residents in care homes. While leveraging skeletons helps to get rid of facial identity and appearance-based information, it is important to ask the question if skeletons can be considered private enough [101], [102] and what steps can be taken to further anonymize the skeletons. Another potential area of investigation for real-world deployment of privacy-protecting anomaly detection systems would be to perform video data acquisition, skeletal tracking (e.g., MediaPipe [103]) and model inferencing in real-time. However, there may be challenges around integrating cloud services, on-chip embedding of AI algorithms, the latency of reaction time, internet stability and false positive rates.

VI. CONCLUSION

In this paper, we provided a survey of recent works that leverage the skeletons or body joints estimated from videos for the anomaly detection task. The skeletons hide the facial identity and overall appearance of people and can provide vital information about joint angles [104], speed of walking [105], and interaction with other people in the scene [17]. Our literature review showed that many deep learning-based approaches leverage reconstruction, prediction error and their other combinations to successfully detect anomalies in a privacy protecting manner. This review suggests the first steps towards increasing adoption of devices (and algorithms) focused on improving privacy in a residential or communal setting. It will further improve the deployment of anomaly detection systems to improve the safety and care of people. The skeleton-based anomaly detection methods can be used to design privacy-preserving technologies for the assisted living of older adults in a care environment [106] or enable older adults to live independently in their own homes to cope with the increasing cost of long-term care demands [107]. Privacy-preserving methods using skeleton features can be employed to assist with skeleton-based rehab exercise monitoring [108] or in social robots for robot-human interaction [109] that assist older people in their activities of daily living.

REFERENCES

[1] S. S. Khan and M. G. Madden, "One-class classification: Taxonomy of study and review of techniques," *Knowl. Eng. Rev.*, vol. 29, no. 3, pp. 345–374, 2014.

[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, 2009.

[3] C. Gautam et al., "Minimum variance-embedded deep Kernel regularized least squares method for one-class classification and its applications to biomedical data," *Neural Netw.*, vol. 123, pp. 191–216, 2020.

[4] P. K. Mishra, C. Gautam, and A. Tiwari, "Minimum variance embedded auto-associative kernel extreme learning machine for one-class classification," *Neural Comput. Appl.*, vol. 33, no. 19, pp. 12973–12987, 2021.

[5] J. Nogas, S. S. Khan, and A. Mihailidis, "Deepfall: Non-invasive fall detection with deep spatio-temporal convolutional autoencoders," *J. Healthcare Informat. Res.*, vol. 4, no. 1, pp. 50–70, 2020.

[6] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, Jan. 2014.

[7] K. Boekhoudt, A. Matei, M. Aghaei, and E. Talavera, "Hr-crime: Human-related anomaly detection in surveillance videos," in *Proc. Int. Conf. Comput. Anal. Images Patterns*, 2021, pp. 164–174.

[8] A. Senior, *Protecting Privacy in Video Surveillance*, vol. 1. Berlin, Germany: Springer, 2009.

[9] P. Climent-Pérez and F. Florez-Revuelta, "Protection of visual privacy in videos acquired with RGB cameras for active and assisted living applications," *Multimedia Tools Appl.*, vol. 80, no. 15, pp. 23649–23664, 2021.

[10] B. Ye et al., "Challenges in collecting Big Data in a clinical environment with vulnerable population: Lessons learned from a study using a multi-modal sensors platform," *Sci. Eng. Ethics*, vol. 25, no. 5, pp. 1447–1466, 2019.

[11] P. Schneider, J. Rambach, B. Mirbach, and D. Stricker, "Unsupervised anomaly detection from time-of-flight depth images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 231–240.

[12] V. Mehta, A. Dhall, S. Pal, and S. S. Khan, "Motion and region aware adversarial learning for fall detection with thermal imaging," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 6321–6328.

[13] S. Denkovski, S. S. Khan, B. Malamis, S. Y. Moon, B. Ye, and A. Mihailidis, "Multi visual modality fall detection dataset," *IEEE Access*, vol. 10, pp. 106422–106435, 2022.

[14] O. Kopuklu, J. Zheng, H. Xu, and G. Rigoll, "Driver anomaly detection: A dataset and contrastive learning approach," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 91–100.

[15] T. Golda, D. Guaia, and V. Wagner-Hartl, "Perception of risks and usefulness of smart video surveillance systems," *Appl. Sci.*, vol. 12, no. 20, 2022, Art. no. 10435.

[16] W. Luo, W. Liu, and S. Gao, "Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection," *Neurocomputing*, vol. 444, pp. 332–337, 2021.

[17] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, "Learning regularity in skeleton trajectories for anomaly detection in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11996–12004.

[18] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: Emotiw 2015," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2015, pp. 423–426.

[19] B. Taati et al., "Algorithmic bias in clinical populations–evaluating and improving facial analysis technology in older adults with dementia," *IEEE Access*, vol. 7, pp. 25527–25534, 2019.

[20] G. Menchetti, Z. Chen, D. J. Wilkie, R. Ansari, Y. Yardimci, and A. E. Çetin, "Pain detection from facial videos using two-stage deep learning," in *Proc. IEEE Glob. Conf. Signal Inf. Process.*, 2019, pp. 1–5.

[21] X. Chen, J. Cheng, R. Song, Y. Liu, R. Ward, and Z. J. Wang, "Video-based heart rate measurement: Recent advances and future prospects," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 10, pp. 3600–3615, Oct. 2019.

[22] T. Gatt, D. Seychell, and A. Dingli, "Detecting human abnormal behaviour through a video generated model," in *Proc. IEEE 11th Int. Symp. Image Signal Process. Anal.*, 2019, pp. 264–270.

[23] O. Temuroglu et al., "Occlusion-aware skeleton trajectory representation for abnormal behavior detection," in *Proc. Int. Workshop Front. Comput. Vis.*, 2020, pp. 108–121.

[24] S. Suzuki, Y. Amemiya, and M. Sato, "Skeleton-based visualization of poor body movements in a child's gross-motor assessment using convolutional auto-encoder," in *Proc. IEEE Int. Conf. Mechatronics*, 2021, pp. 1–6.

[25] Z. Jiang, G. Song, Y. Qian, and Y. Wang, "A deep learning framework for detecting and localizing abnormal pedestrian behaviors at grade crossings," *Neural Comput. Appl.*, vol. 34, no. 24, pp. 22099–22113, 2022.

[26] G. Song, Y. Qian, and Y. Wang, "Analysis of abnormal pedestrian behaviors at grade crossings based on semi-supervised generative adversarial networks," *Appl. Intell.*, vol. 53, no. 19, Oct. 2023.

[27] Z. Fan, S. Yi, D. Wu, Y. Song, M. Cui, and Z. Liu, "Video anomaly detection using cyclegan based on skeleton features," *J. Vis. Commun. Image Representation*, vol. 85, 2022, Art. no. 103508.

[28] R. Rodrigues, N. Bhargava, R. Velmurugan, and S. Chaudhuri, "Multi-timescale trajectory prediction for abnormal human activity detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2020, pp. 2626–2634.

[29] X. Zeng, Y. Jiang, W. Ding, H. Li, Y. Hao, and Z. Qiu, "A hierarchical spatio-temporal graph convolutional neural network for anomaly detection in videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 200–212, Jan. 2023.

[30] B. Fan, P. Li, S. Jin, and Z. Wang, "Anomaly detection based on pose estimation and GRU-FFN," in *Proc. IEEE Sustain. Power Energy Conf.*, 2021, pp. 3821–3825.

[31] W. Pang, Q. He, and Y. Li, "Predicting skeleton trajectories using a skeleton-transformer for video anomaly detection," *Multimedia Syst.*, vol. 28, no. 4, pp. 1481–1494, 2022.

[32] C. Huang et al., "Hierarchical graph embedded pose regularity learning via spatio-temporal transformer for abnormal behavior detection," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 307–315.

[33] Y. Li and Z. Zhang, "Video abnormal behavior detection based on human skeletal information and GRU," in *Proc. Int. Conf. Intell. Robot. Appl.*, 2022, pp. 450–458.

[34] N. Li, F. Chang, and C. Liu, "Human-related anomalous event detection via spatial-temporal graph convolutional autoencoder with embedded long short-term memory network," *Neurocomputing*, vol. 490, pp. 482–494, 2021.

[35] T.-H. Wu, C.-L. Yang, L.-L. Chiu, T.-W. Wang, G. J. Faure, and S.-H. Lai, "Confidence-aware anomaly detection in human actions," in *Proc. Asian Conf. Pattern Recognit.*, 2022, pp. 240–254.

[36] S. Luo, S. Wang, Y. Wu, and C. Jin, "Memory enhanced spatial-temporal graph convolutional autoencoder for human-related video anomaly detection," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis.*, 2022, pp. 665–677.

[37] N. Li, F. Chang, and C. Liu, "Human-related anomalous event detection via memory-augmented wasserstein generative adversarial network with gradient penalty," *Pattern Recognit.*, vol. 138, 2023, Art. no. 109398.

[38] A. Markovitz, G. Sharir, I. Friedman, L. Zelnik-Manor, and S. Avidan, "Graph embedded pose clustering for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10539–10547.

[39] T. Cui, W. Song, G. An, and Q. Ruan, "Prototype generation based shift graph convolutional network for semi-supervised anomaly detection," in *Proc. Chin. Conf. Image Graph. Technol.*, 2021, pp. 159–169.

[40] C. Liu, R. Fu, Y. Li, Y. Gao, L. Shi, and W. Li, "A self-attention augmented graph convolutional clustering networks for skeleton-based video anomaly behavior detection," *Appl. Sci.*, vol. 12, no. 1, 2022, Art. no. 4.

[41] X. Chen, S. Kan, F. Zhang, Y. Cen, L. Zhang, and D. Zhang, "Multiscale spatial temporal attention graph convolution network for skeleton-based anomaly behavior detection," *J. Vis. Commun. Image Representation*, vol. 90, 2023, Art. no. 103707.

[42] M. Yan, Y. Xiong, and J. She, "Memory clustering autoencoder method for human action anomaly detection on surveillance camera video," *IEEE Sensors J.*, vol. 23, no. 18, pp. 20715–20728, Sep. 2023.

[43] Y. Yang, Z. Fu, and S. M. Naqvi, "A two-stream information fusion approach to abnormal event detection in video," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 5787–5791.

[44] M. H. Javed, Z. Yu, T. Li, N. Anwar, and T. M. Rajeh, "learning anomalous human actions using frames of interest and decoderless deep embedded clustering," in *Proc. Int. J. Mach. Learn. Cybern.*, 2023, pp. 1–15.

[45] N. Li, F. Chang, and C. Liu, "A self-trained spatial graph convolutional network for unsupervised human-related anomalous event detection in complex scenes," *IEEE Trans. Cogn. Devel. Syst.*, vol. 15, no. 2, pp. 737–750, Jun. 2023.

[46] H. Tani and T. Shibata, "Frame-wise action recognition training framework for skeleton-based anomaly behavior detection," in *Proc. Int. Conf. Image Anal. Process.*, 2022, pp. 312–323.

[47] F. Sato, R. Hachiuma, and T. Sekii, "Prompt-guided zero-shot anomaly action recognition using pretrained deep skeleton features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 6471–6480.

[48] L. Song, G. Yu, J. Yuan, and Z. Liu, "Human pose estimation and its application to action recognition: A survey," *J. Vis. Commun. Image Representation*, vol. 76, 2021, Art. no. 103055.

[49] A. Badiola-Bengoa and A. Mendez-Zorrilla, "A systematic review of the application of camera-based human pose estimation in the field of sport and physical exercise," *Sensors*, vol. 21, no. 18, 2021, Art. no. 5996.

[50] K. Boekhoudt and E. Talavera, "Spatial-temporal transformer for crime recognition in surveillance videos," in *Proc. IEEE 18th Int. Conf. Adv. Video Signal Based Surveill.*, 2022, pp. 1–8.

[51] W. Du, Y. Wang, and Y. Qiao, "Rpan: An end-to-end recurrent pose-attention network for action recognition in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3725–3734.

[52] S. Suzuki, Y. Amemiya, and M. Sato, "Enhancement of child gross-motor action recognition by motional time-series images conversion," in *Proc. IEEE/SICE Int. Symp. Syst. Integration*, 2020, pp. 225–230.

[53] D. Gong et al., "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1705–1714.

[54] Y. Tang, L. Zhao, S. Zhang, C. Gong, G. Li, and J. Yang, "Integrating prediction and reconstruction for anomaly detection," *Pattern Recognit. Lett.*, vol. 129, pp. 123–130, 2020.

[55] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 341–349.

[56] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb d: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1010–1019.

[57] W. Kay et al., "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.

[58] S. S. Khan and A. Ahmad, "Cluster center initialization algorithm for k-means clustering," *Pattern Recognit. Lett.*, vol. 25, no. 11, pp. 1293–1302, 2004.

[59] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, *arXiv:1804.02767*.

[60] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discov. From Data*, vol. 6, no. 1, pp. 1–39, 2012.

[61] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12026–12035.

[62] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, vol. 32, no. 1 2018, doi: 10.1609/aaai.v32i1.12328.

[63] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.

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

[65] D. M. Blei and M. I. Jordan, "Variational inference for dirichlet process mixtures," *Bayesian Anal.*, vol. 1, no. 1, pp. 121–144, 2006.

[66] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in matlab," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2720–2727.

[67] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2018, pp. 6479–6488.

[68] University of Minnesota, "Monitoring human activity." [Online]. Available: https://mha.cs.umn.edu/proj_events.shtml

[69] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 168–172.

[70] M. Cheng, K. Cai, and M. Li, "RWF-2000: An open large scale video database for violence detection," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 4183–4190.

[71] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2334–2343.

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

[73] G. Papandreou et al., "Towards accurate multi-person pose estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4903–4911.

[74] T. Sekii, "Pose proposal networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 342–357.

[75] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5693–5703.

[76] L. Ruff et al., "A unifying review of deep and shallow anomaly detection," *Proc. IEEE*, vol. 109, no. 5, pp. 756–795, May 2021.

[77] S. S. Khan and J. Hoey, "Dtfall: Decision-theoretic framework to report unseen falls," in *Proc. PervasiveHealth*, 2016, pp. 146–153.

[78] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS One*, vol. 10, no. 3, 2015, Art. no. e0118432.

[79] Y. M. Galvão, L. Portela, J. Ferreira, P. Barros, O. A. D. A. Fagundes, and B. J. Fernandes, "A framework for anomaly identification applied on fall detection," *IEEE Access*, vol. 9, pp. 77264–77274, 2021.

[80] S. S. Khan et al., "Unsupervised deep learning to detect agitation from videos in people with dementia," *IEEE Access*, vol. 10, pp. 10349–10358, 2022.

[81] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *Comput. Vis. Image Understanding*, vol. 192, 2020, Art. no. 102897.

[82] Y. Cheng, B. Yang, B. Wang, W. Yan, and R. T. Tan, "Occlusion-aware networks for 3D human pose estimation in video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 723–732.

[83] S. Gong, T. Xiang, and S. Hongeng, "Learning human pose in crowd," in *Proc. 1st ACM Int. Workshop Multimodal Pervasive Video Anal.*, 2010, pp. 47–52.

[84] U. Iqbal and J. Gall, "Multi-person pose estimation with local joint-to-person associations," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 627–642.

[85] S. S. Khan, M. E. Karg, D. Kulić, and J. Hoey, "Detecting falls with x-factor hidden Markov models," *Appl. Soft Comput.*, vol. 55, pp. 168–177, 2017.

[86] S. A. Boamah, R. Weldrick, T.-S. J. Lee, and N. Taylor, "Social isolation among older adults in long-term care: A scoping review," *J. Aging Health*, vol. 33, no. 7/8, pp. 618–632, 2021.

[87] T.-N. Nguyen, H.-H. Huynh, and J. Meunier, "Skeleton-based abnormal gait detection," *Sensors*, vol. 16, no. 11, 2016, Art. no. 1792.

[88] R. Baptista, G. Demisse, D. Aouada, and B. Ottersten, "Deformation-based abnormal motion detection using 3D skeletons," in *Proc. 8th Int. Conf. Image Process. Theory, Tools Appl.*, 2018, pp. 1–6.

[89] T. Warren, "Microsoft kills off Kinect, stops manufacturing it," 2017, Accessed: Feb. 23, 2022. [Online]. Available: https://www.theverge.com/2017/10/25/16542870/microsoft-kinect-dead-stop-manufacturing

[90] "Vicon," 2019. [Online]. Available: https://www.vicon.com/

[91] "Thecaptury," 2019. [Online]. Available: https://thecaptury.com/

[92] AltumView, "Sentinare 2," 2022, Accessed: Feb. 24, 2022. [Online]. Available: https://altumview.ca/

[93] S. S. Khan, P. K. Mishra, B. Ye, K. Newman, A. Iaboni, and A. Mihailidis, "Empirical thresholding on spatio-temporal autoencoders trained on surveillance videos in a dementia care unit," in *Proc. Conf. Robots Vis.*, 2023, pp. 265–272.

[94] E. Duman and O. A. Erdem, "Anomaly detection in videos using optical flow and convolutional autoencoder," *IEEE Access*, vol. 7, pp. 183914–183923, 2019.

[95] A. Abedi and S. S. Khan, "Fedsl: Federated split learning on distributed sequential data in recurrent neural networks," *Multimedia Tools and Appl.*, Sep. 2023, doi: 10.1007/s11042-023-15184-5.

[96] J. Yan, F. Angelini, and S. M. Naqvi, "Image segmentation based privacy-preserving human action recognition for anomaly detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 8931–8935.

[97] P. K. Mishra, A. Iaboni, B. Ye, K. Newman, A. Mihailidis, and S. S. Khan, "Privacy-protecting behaviours of risk detection in people with dementia using videos," *BioMedical Eng. OnLine*, vol. 22, no. 1, pp. 1–17, 2023.

[98] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3D human pose estimation with spatial and temporal transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11656–11665.

[99] C.-L. Zhang, J. Wu, and Y. Li, "Actionformer: Localizing moments of actions with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 492–510.

[100] W. Feng, R. Liu, and M. Zhu, "Fall detection for elderly person care in a vision-based home surveillance environment using a monocular camera," *Signal, Image Video Process.*, vol. 8, no. 6, pp. 1129–1138, 2014.

[101] H. Wang and L. Wang, "Learning content and style: Joint action recognition and person identification from human skeletons," *Pattern Recognit.*, vol. 81, pp. 23–35, 2018.

[102] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognit.*, vol. 98, 2020, Art. no. 107069.

[103] C. Lugaresi et al., "Mediapipe: A framework for building perception pipelines," 2019, *arXiv:1906.08172*.

[104] Q. Guo and S. S. Khan, "Exercise-specific feature extraction approach for assessing physical rehabilitation," in *Proc. 4th IJCAI Workshop AI Aging, Rehabil. Intell. Assist. Living*, 2021.

[105] J. Kovač and P. Peer, "Human skeleton model based dynamic features for walking speed invariant gait recognition," *Math. Problems Eng.*, vol. 2014, Jan.21, 2014, doi: 10.1155/2014/484320..

[106] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "A review on vision techniques applied to human behaviour analysis for ambient-assisted living," *Expert Syst. With Appl.*, vol. 39, no. 12, pp. 10873–10888, 2012.

[107] Y. Hbali, S. Hbali, L. Ballihi, and M. Sadgal, "Skeleton-based human activity recognition for elderly monitoring systems," *IET Comput. Vis.*, vol. 12, no. 1, pp. 16–26, 2018.

[108] Š. Obdržálek et al., "Accuracy and robustness of kinect pose estimation in the context of coaching of elderly population," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2012, pp. 1188–1193.

[109] M. Garcia-Salguero, J. Gonzalez-Jimenez, and F.-A. Moreno, "Human 3D pose estimation with a tilting camera for social mobile robot interaction," *Sensors*, vol. 19, no. 22, 2019, Art. no. 4943.

Alex Mihailidis received the Ph.D. and P.Eng. degrees. He is currently the Barbara G. Stymiest Research Chair of rehabilitation technology, KITE Research Institute, University Health Network/University of Toronto, Toronto, ON, Canada. He is also the Scientific Director of the AGE-WELL Network of Centres of Excellence, which focuses on the development of new technologies and services for older adults. He is a Professor with the Department of Occupational Science and Occupational Therapy and with the Institute of Biomedical Engineering, University of Toronto (U of T), and holds a cross appointment with the Department of Computer Science, U of T. He is very active in the rehabilitation engineering profession and the Immediate Past President of the Rehabilitation Engineering and Assistive Technology Society for North America (RESNA). He was named a Fellow of RESNA in 2014, which is one of the highest honours within this field of research and practice. He has authored or coauthored more than 150 journal and conference papers in this field and co-edited two books: *Pervasive computing in Healthcare* and *Technology and Aging*. His research interests include biomedical and biochemical engineering, computer science, geriatrics and occupational therapy. He is an internationally recognized Researcher in the field of technology and aging.

Shehroz S. Khan received the B.Sc. Eng. degree, the master's and Ph.D. degrees in computer science in 1997, 2010, and 2016, respectively. He is currently a Scientist with KITE – Toronto Rehabilitation Institute (TRI), University Health Network, Toronto, ON, Canada. He is also cross appointed as an Assistant Professor with the Institute of Biomedical Engineering, University of Toronto (UofT), Toronto. He was a Postdoctoral Researcher with the UofT and TRI. Prior to joining academics, he worked in various scientific and Researcher roles in the industry and government jobs. He has authored or coauthored 49 peer-reviewed research papers. His research focuses on the development of AI algorithms for solving aging related health problems. He has organized four editions of the peer-reviewed workshop on AI in Aging, Rehabilitation and Intelligent Assisted Living held with top AI conferences (ICDM and IJCAI) from 2017 to 2021. His research is funded through several granting agencies in Canada and abroad, including NSERC, CIHR, AGEWELL, SSHRC, CABHI, AMS Healthcare, JP Bickell Foundation, United Arab Emirates University and LG Electronics. He is an Associate Editor for the *Journal of Rehabilitation and Assistive Technologies*.

Pratik K. Mishra received the master's degree in computer science and engineering from the Indian Institute of Technology (IIT) Indore, Indore, India, in 2020. He is currently working toward the Ph.D. degree with the Institute of Biomedical Engineering, University of Toronto. He is working on the application of computer vision for detecting behaviours of risk in people with dementia. He was a research Volunteer with the Toronto Rehabilitation Institute, Toronto, ON, Canada, and as a Data Management Support Specialist with IBM India Private Limited.