Deep Learning Methods for Single Camera Based Clinical In-bed Movement Action Recognition*

Tamás Karácsony[a,b,c,*], László Attila Jeni[c], Fernando De la Torre[c] and João Paulo Silva Cunha[a,b,**]

[a]Center for Biomedical Engineering Research, Institute for Systems' Engineering and Computers, Technology & Science (INESC TEC), Porto, Portugal
[b]Faculty of Engineering (FEUP), University of Porto, Porto, Portugal
[c]Robotics Institute at Carnegie Mellon University, Pittsburgh, 15213, PA, USA

ARTICLE INFO

Keywords:
Action recognition
3D Motion Capture
Clinical In-bed monitoring
Diagnosis support
Seizure semiology
Epilepsy

ABSTRACT

Many clinical applications involve in-bed patient activity monitoring, from intensive care and neuro-critical infirmary, to semiology-based epileptic seizure diagnosis support or sleep monitoring at home, which require accurate recognition of in-bed movement actions from video streams.

The major challenges of clinical application arise from the domain gap between common in-the-lab and clinical scenery (e.g. viewpoint, occlusions, out-of-domain actions), the requirement of minimally intrusive monitoring to already existing clinical practices (e.g. non-contact monitoring), and the significantly limited amount of labeled clinical action data available.

Focusing on one of the most demanding in-bed clinical scenarios - semiology-based epileptic seizure classification – this review explores the challenges of video-based clinical in-bed monitoring, reviews video-based action recognition trends, monocular 3D MoCap, and semiology-based automated seizure classification approaches. Moreover, provides a guideline to take full advantage of transfer learning for in-bed action recognition for quantified, evidence-based clinical diagnosis support.

The review suggests that an approach based on 3D MoCap and skeleton-based action recognition, strongly relying on transfer learning, could be advantageous for these clinical in-bed action recognition problems. However, these still face several challenges, such as spatio-temporal stability, occlusion handling, and robustness before realizing the full potential of this technology for routine clinical usage.

1. Introduction

Clinical in-bed patient movement monitoring is crucial for several aspects of disease management in many healthcare applications, such as intensive care and neuro-critical infirmary, semiology-based epileptic seizure diagnosis support, or sleep monitoring at home. In several of these fields, it is essential to provide quantified movement analysis, in order to provide evidence-based diagnosis support for clinicians. For example, in the case of epilepsy, this includes the movement quantification of seizure semiology, advancing from the current clinical practice, which is a qualitative visual inspection of seizure videos, searching for Movement of Interests (MOIs), to an automated evidence-based approach [110, 111, 132, 64, 66, 65].

The following review is presented from the perspective of one of the most challenging applications of clinical in-bed action recognition; epileptic seizure classification based on seizure semiology in Epilepsy Monitoring Units (EMUs) (Fig. 1). This sub-field includes all of the major clinical challenges and represents well the general principles, remaining challenges, and possible solutions for video-based clinical in-bed action recognition applications in general.

Although there are several approaches to monitor movements, such as Inertial Measurement Units (IMUs), pressure sensors embedded in the bed, and marker-based Motion Capture (MoCap), it is not desirable to attach any sensor or marker to the patient, as they can be displaced or detached (for example during violent seizures), and these solutions are rather uncomfortable for the patient. They interfere with common clinical practice, therefore the most promising approaches are the markerless, computer vision (CV) technologies, specifically, Deep Learning (DL) based techniques, in terms of performance these approaches dominate the CV field, due to recent developments in the field. These can include processing data streams of RGB, Infrared (IR), and depth videos. For quantitative semiology characterization, automated and semi-automated computer-vision analysis approaches have been a promising methodology [43], but still depend on considerable human interaction [15]. In EMUs IR video streams ensure nightly monitoring, without RGB availability due to ambient lighting [52, 10, 68, 77]. Furthermore, these movements can be classified as seizures or MOIs, which can be utilized as a diagnosis support

*This work was partially funded by Fundação para a Ciência e a Tecnologia under the scope of the CMU Portugal program (Ref PRT/BD/152202/2021). This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020. DOI 10.54499/LA/P/0063/2020 | https://doi.org/10.54499/LA/P/0063/2020

*Corresponding author
**Principal corresponding author
tamas.karacsony@inesctec.pt (T. Karácsony); jcunha@ieee.org (J.P.S. Cunha)
https://tamaskaracsony.github.io/ (T. Karácsony); https://www.laszlojeni.com/ (L.A. Jeni); https://www.cs.cmu.edu/~ftorre/ (F.D.l. Torre); https://www.inesctec.pt/en/people/joao-paulo-cunha (J.P.S. Cunha)
ORCID(s): 0000-0002-7899-1786 (T. Karácsony); 0000-0002-2830-700X (L.A. Jeni); 0000-0003-4131-9045 (J.P.S. Cunha)

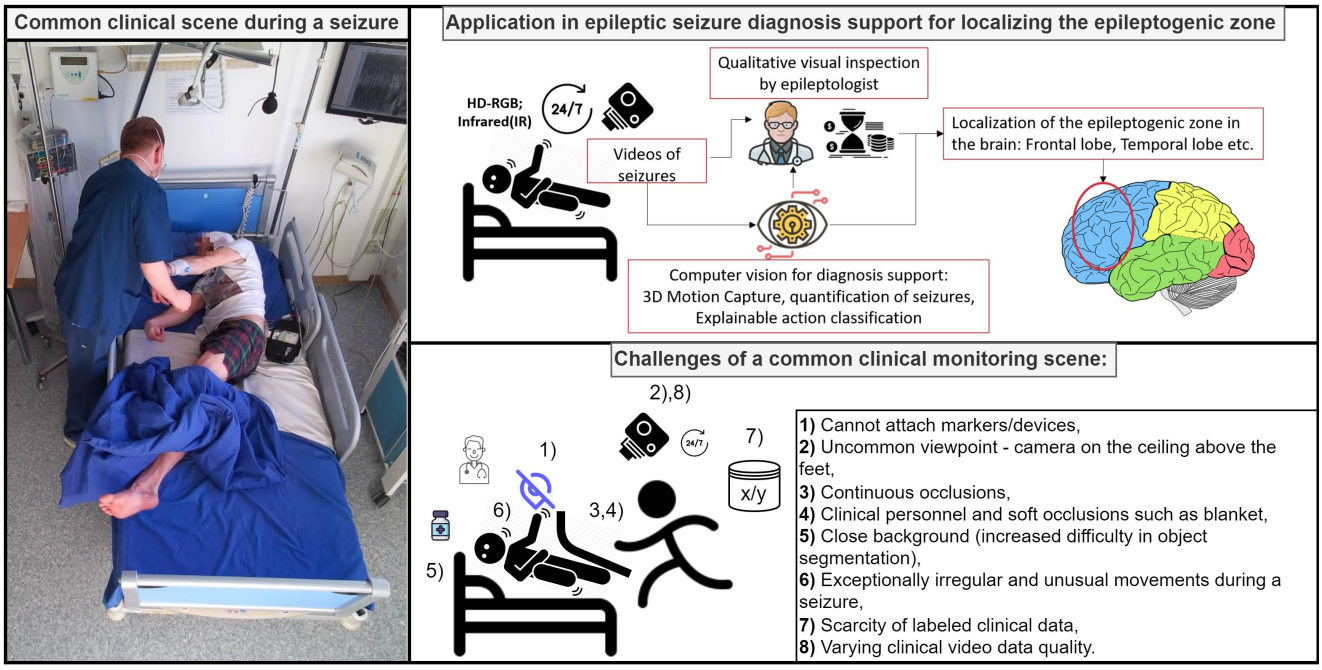

Figure 1: A complex application of in-bed movement action recognition demonstrated through diagnosis support of epileptic seizure classification in an Epilepsy Monitoring Unit (EMU) environment, highlighting the common challenges of in-bed action recognition of this in-bed patient monitoring scenarios that are common to many others.

tool, alarms, and prediction of seizures in clinics. The current state-of-the-art (SOTA), end-to-end DL approaches for epileptic seizure classification are promising; however, they are not yet explainable, and can not be utilized for movement quantification [64, 66], or they suffer from significant noise from the background, modest performance or overfitting to subject-specific features [2, 3, 5, 7, 6, 4].

In summary, computer vision-based markerless approaches are ideal approaches for clinical movement-based diagnosis support, which can be utilized for movement quantification, classification, and quantified disease analysis. However they require significant amounts of research before overcoming their limitations and including them in clinical practice.

The contributions of this paper are the following:

- Identifying clinical challenges of video-based clinical in-bed action recognition and MoCap
- Provide a review of video-based motion capture and action recognition from a clinical in-bed monitoring perspective
- Provide a review of state-of-the-art semiology-based epileptic seizure classification approaches from videos
- Discusses challenges and possible solutions and determines that a 2-stage 3D MoCap and action recognition approach is a promising future research direction

2. Challenges of video-based clinical in-bed movement classification and MoCap

There are two main directions to classify actions from the movements of a person, either utilizing the raw video and classifying the action in an end-to-end fashion [19, 72, 172, 175], or utilizing keypoint, skeleton feature extraction as a first stage and then carry out classification on the skeletons [99, 27, 142].

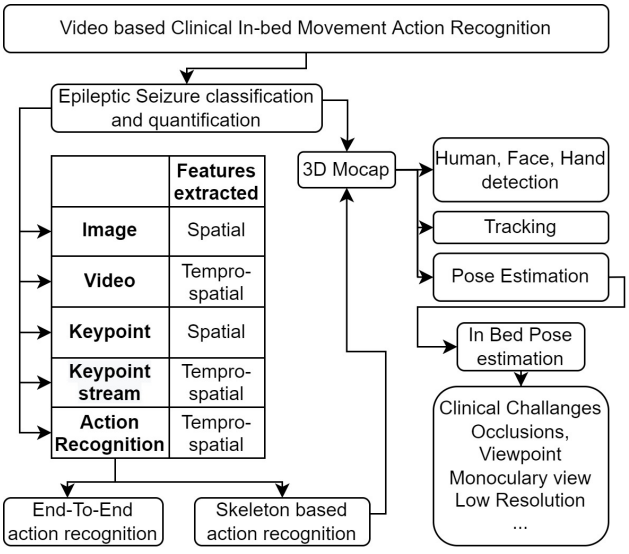

Figure 2: Overview of the topics covered by this review and their relations

Due to the clinical non-contact monitoring requirement only marker-less motion capture approaches are considered, which extract from the raw video the skeleton keypoints of the person during the movement. To achieve a viewpoint invariant representation for the later action recognition stage, achieving 3D MoCap is preferred to only 2D approaches. Furthermore, due to the limited clinical action training data it is advantageous to solve the challenge of viewpoint invariance during MoCap, where general datasets can be utilized for training, as opposed to utilizing scarce clinical data to learn viewpoint invariance, as detailed later. Thus, in the following, we cover the review of the challenges of 3D MoCap in this scenario, the full overview of the topics covered by this review and their relations are summarized on Fig. 2.

These markerless MoCap systems include, single monocular RGB [131, 119, 167] RGB-D [161], only depth [176, 179], image or video data-based approaches to mention a few. These perform well in controlled scenarios, with sufficient lightning, HD RGB input, and with minimal occlusions. However, they do not or only partially address the clinical MoCap challenges originating from 24/7 in-bed monitoring. These challenges are the following: 1) Due the clinical practice markers can not be attached to the patient, 2) uncommon viewpoint, 3) the presence of continuous occlusions, 4) clinical personnel and soft occlusions such as blanket, 5) close background, 6) exceptionally irregular and unusual movements during seizures, 7) scarcity of labeled clinical data 8) and the varying clinical video data quality (Fig. 1).

Here the uncommon viewpoint refers both to people lying in bed as opposed to the standing/sitting scenario, which builds up the majority of the relevant datasets, and the view angle related to the person. The common camera viewpoint of available 3D MoCap datasets is pointing from the frontal direction from around eye height, usually ranging from around hip height to above the head point of view, however in this scenario as the camera is hung from the ceiling from below the plane of the feet (Fig. 1), practically translating this to the standing person scenario this would refer to placing a camera to the ground level and even below ground level. Both of these are quite rarely represented in datasets, leading to a slightly out-of-distribution scenario, which might impact the performance of the architectures. Moreover, segmenting subjects with a close background presents difficulties as the person in bed may blend seamlessly into their surroundings. Traditional depth-based segmentation methods, which typically employ point cloud-based semantic segmentation or depth thresholding, face major challenges in these settings. The primary issue arises from the continuous surface where the subject is in contact with the background. Likewise, RGB segmentation encounters some obstacles in clearly defining boundaries when the subject is in contact and surrounded close by a soft fabric as a background, often leading to minor but impactful occlusions around the edges, breaking the smooth continuity of the boundaries. Additionally, due to the 24/7 h monitoring requirement, at night only IR B/W, sometimes with low-resolution and depth videos may be available.

Although occlusions can be partially addressed with multi-camera systems, and tracking performance can be improved [128]. However, in clinical practice, the system has to be cost-efficient, occupy minimal space, and has to store monitoring data for several days on a reasonable PC. Moreover, clinical monitoring rooms are commonly optimized to fit the maximum number of patients and are cluttered with clinical equipment, thus the available space and positions to mount cameras can be significantly restricted. These parameters point to limiting the number of used cameras, therefore the most widespread solution is to utilize a single camera for monitoring. Considering the large data requirement for DL classification of clinical actions such as MOIs and seizures the approach has to be scalable on already existing clinical practices and data. Thus the research will focus on monocular monitoring approaches.

In summary, video-based clinical in-bed action recognition and 3D MoCap, also commonly referred to in the literature as 3D Human Pose Estimation (HPE) and tracking, have to be adopted for the clinical scenarios, integrated and improved to overcome the challenges of clinical applications.

3. Deep learning for video-based action recognition

Human actions can be represented using various data modalities, such as RGB, skeleton, depth, infrared, point cloud, event stream, audio, acceleration, radar, and even WiFi signals [147]. We kindly refer the reader to an extensive review for the general details of all these modalities used in human action recognition and their benefits and disadvantages to [147], moreover a more general outlook is provided in [117].

In clinical action monitoring, such as epileptic seizure monitoring, the widespread monitoring infrastructure consists of clinical-grade monocular RGB and IR IP cameras compatible with other clinical systems. Thus, it is advantageous to consider this existing infrastructure for the action recognition approaches, as clinics already store these historical clinical RGB and IR data, which can be vital for up-scaling data collection for training of DL approaches to classify clinical actions. Therefore, these two raw modalities are the most favorable.

Infrared data can be utilized with similar approaches for action recognition as RGB, as described in detail in [147]. It is optimal for night monitoring, however, faces challenges, like low contrast and poor signal-to-noise ratio, making robust action recognition difficult in standard and even more so in clinical settings. As thoroughly analyzed in [147], depth data or point clouds could also be viable options. Yet, their drawbacks, such as significant computational complexity, extensive data storage needs, and the absence of pre-existing clinical data, infrastructure, and the lack of color and texture,

combined with their typically limited resolution, frame rate, and effective range, make them less desirable approaches.

As stated before for clinical applications in EMUs, such as for epileptic seizure classification, the 3D skeleton, MoCap-based approaches are the most advantageous. Besides the inherent quantification of movements and some degree of explainability of this approach, 3D skeleton-based human action recognition (HAR) is insensitive to the viewpoint and background of the videos. Therefore, already available datasets, which commonly represent standing people from a frontal view, may still be utilized for pre-training the action classification network in a transfer learning approach. Additionally, the lack of appearance and detailed shape information is an advantage for this application, as it promotes generalizability across patients.

RGB and skeleton human action recognition approaches commonly utilize the NTU RGB+D [137], NTU RGB+D 120 [93], Kinetics-400, Kinetics-600, Kinetics-700 [67, 18, 141], UCF101 [143], and HMDB51 [73] datasets and their derivatives for training and evaluation.

Both RGB and IR modalities, commonly recorded in clinical settings, can be employed in two primary ways: through end-to-end processing or by extracting skeletons for action recognition. Our goal here is to outline the principal methods in both end-to-end and skeleton-based techniques. With a significantly larger body of research available on RGB-based end-to-end methods, our focus will predominantly be on comparing RGB and skeleton-based approaches. While IR-based methods adhere to similar principles, they face unique challenges due to smaller and less frequent training datasets. Often, they rely on common RGB-based datasets and must address the domain shift issues we previously discussed.

In the following sections, state-of-the-art RGB-based end-to-end approaches, as the most widely utilized approaches, and skeleton-based methods are reviewed.

3.1. Monocular RGB end-to-end approaches

The RGB-based approaches provide rich appearance information, are easy to obtain and operate, and have a wide range of operations; however, they are sensitive to the viewpoint, background, and illumination [147]. Moreover, this representation stores the spatio-temporal features representing the actions with massive data sizes, which lead to large computational costs to capture these connections [147].

These approaches can be categorized as multi-stream 2D CNN, RNN, 3D CNN-based, and transformer-based methods and combinations of these. The 2D CNN-based methods generally are very limited in capturing the long-term dependencies, therefore 2D CNN-s were combined with LSTMs to capture the long-term connections [147]. In order to improve the capture of local and global spatio-temporal connection of the actions through the videos 3D CNNs were utilized, such as I3D [19]. To further enhance the capture of long-term connections a 3D CNN-LSTM architecture was proposed [162]. Recently, MoViNets, an efficient 3D CNN approach was proposed [72], which has state-of-the-art performance, among 3D CNN based approaches, on several benchmark datasets [72]. It utilizes the 3D CNNs with stream buffers and temporal ensembles, which enables the architecture to be online inferred with a fraction of the memory and computational footprint, than other state-of-the-art methods with retaining state-of-the-art performance. Recently, transformer-based architectures further advanced the state-of-the-art previously set by 3D CNNs, achieving significant performance improvements [172, 175]. These RGB video-based end-to-end approaches establish a strong baseline for action recognition.

3.2. Skeleton-based approaches

Skeleton-based action recognition approaches have received increasing attention in recent years, as a compact, but an informative representation of human actions. It is insensitive to viewpoint, background, and provides the 3D trajectories of human motion. However, it lacks of appearance and detailed shape information and can be a noisy representation [147]. These later properties may be disadvantageous in some HAR cases, where the surrounding scene or appearance may have additional features to improve HAR performance. Nevertheless, in the case of seizure classification these properties are advantages, as these mitigate any overfit originating from the patient, background, or scenery-related features. In contrast, RGB video-based approaches may include these seizure unrelated features.

Early approaches of skeleton-based HAR utilized RNNs and CNNs, which recently were outperformed by Graph Convolutional Networks (GCNs) [147]. Skeleton-based action recognition can naturally be represented in graphs, due to the joint dependency structure, therefore utilizing GCNs is beneficial. Current top performing GCNs are MS-G3D [99], ST-GCN++ [40], CTR-GCN [27] and EfficientGCN variants [142]. MS-G3D introduced disentangled and unifying graph convolutions to improve spatio-temporal information flow for feature extraction [99]. ST-GCN++ [40] modified the temporal module from a 1D convolution to a multi-branch temporal ConvNet (TCN) and the spatial module joint features fusing strategy of the original Spatial Temporal GCN [173]. EfficientGCN proposed to utilize [142] early fused multiple input branches (relative and absolute joint position, velocity, lengths and angles of bones), spatial-temporal joint attention, moreover to further improve model efficiencies and reduce the model complexity introduced four temporal convolutional layers based on separable convolution [142]. Currently, this approach achieves SOTA performance on NTU RGB+D [137], NTU RGB+D 120 [93], while keeping the network very efficient.

A strategy to train symbiotic GNNs for HAR and motion prediction was proposed [83], these two goals during training can complement each other and improve performance for both [83]. Although GCNs are dominating currently the field, PoseConv3D a CNN-based method was proposed recently utilizing 3D heatmaps volumes [41]. The authors suggested that it is more effective in learning spatiotemporal

General datasets			
Dataset name	Top Model	Note	Data Year
Human3.6M [56]	TesseTrack [128]	Largest base	2014
CMU Panoptic [61]	TesseTrack [128]	10 (RGB-D) + 480 (VGA) + + 30 (HD) camera dome	2016-2019
3DPW [102]	DynaBOA [47]	Best in the wild	2018
MPI-INF-3DHP [104]	SPIN [71]	In & outdoor	2018
HumanEva-I [140]	Lifting Transformer [84]	-	2010
Total Capture [153]	GeoFuse [182]	8 camera, 12 IMU	2017
AGORA [118]	SPEC [70]	Synthetic	2021
Surreal [155]	[165]	Synthetic	2017

(a) General datasets for 3D MoCap

Clinical datasets			
Dataset name	Top Model	Note	Data Year
SLP [96, 95]	[95]	Clinical multimodal Lying Pose	2019
BlanketGen [16]	[16]	Synthetic Blanket occlusion on 3DPW [102]	2022
BlanketSet [17]	[16]	Clinical RGB-D-IR semi-synchronized	2022
MVOR [144]	[62]	Clinical multivew RGB-D	2018
Patient Mocap [1]	[1]	Synthetic Blanket occlusion	2016

(b) Clinically relevant datasets for evaluation of 3D MoCap

Table 1
Current popular and clinically relevant datasets for evaluation of 3D MoCap

features, more robust against pose estimation noises, and generalizes better in cross-dataset settings, moreover, it can handle multiple-person scenarios without additional computation cost [41].

In conclusion, the state-of-the-art suggests that utilizing skeleton data GCN-based action recognition is the most advantageous approach for seizure classification. Although there is a risk of low performance due to noisy skeleton information [147], which strongly depends on the MoCap performance in the clinical environment. In this case, an alternative approach may be a CNN-based approach, such as PoseConv3D [41].

4. Monocular Video Based Clinical 3D Motion Capture

4.1. Datasets

There are a vast amount of datasets available for 3D human MoCap, an extensive review of 171 3D skeleton-based human representations, including 150 papers is presented by Han et. al. [49] represents the SOTA, up until 2017. Most of these datasets however are not utilized for SOTA solutions. Therefore a short summary of current popular datasets for 3D human pose estimation is presented in Tab. 1a.

Datasets from the perspective of clinical patient monitoring are summarized in Tab. 1b. A Multi-view RGB-D operating room (MVOR) dataset was developed to test the visual challenges present in such environments, such as occlusions and clutter. The dataset consists of 732 synchronized multi-view frames recorded by three RGB-D cameras [144]. In this paper, the issue of data privacy was also addressed, which is required for medical confidentiality. In order to publicly release the dataset the authors were obligated to hide the identity and nudity of people, thus blurring some parts of the images, which presents another visual challenge with clinical monitoring. In the presented comparative study of how the baselines have been impacted by the blurring it was concluded that it only affects mildly the accuracy of detections [144].

The Simultaneously-collected multimodal Lying Pose (SLP) dataset [96, 95] was collected with the goal to improve in-bed human pose estimation for clinical applications. It consists of 14.7K images from 109 participants captured using multiple imaging modalities including RGB, long wave infrared (LWIR), depth, and pressure map. The participants lie in bed in different positions and blanket occlusion conditions.

BlanketSet [17] is a real-word clinical action recognition and semi-synchronized MoCap dataset, consisting of 405 RGB-D-IR videos, with 15 participants carrying out 8 different actions, with movements semi-synchronized in different blanket occlusion conditions [17].

BlanketGen [16] a synthetic blanket occlusion augmentation pipeline [16] was proposed and demonstrated on 3DPW [102], generating BlanketGen-3DPW [16]. Another semi-synthetic clinical dataset was proposed, where an occluding blanket was simulated to improve patient MoCap [1]; however, this dataset only utilized depth data and it was not made publicly available.

4.2. SOTA markerless 3D MoCap

An extensive review of 3D human pose estimation algorithms for markerless motion capture is provided in [37], which summarizes the previous surveys on the field, and extends the full review until 2021. Even broader surveys cover both 2D and 3D Human Pose Estimation (HPE) [184, 98], including additionally commonly used evaluation metrics, popular datasets, architectures, and approaches. Here, the SOTA is covered from a clinical patient monitoring perspective of RGB inputs.

4.2.1. Top-down vs bottom-up approaches of 3D MoCap

The Top-Down approach first detects all individual persons on the frame and extracts them, relying on a SOTA object detection network. Then for each person individually estimates the 3D human poses and meshes. This approach can take more advantage of models, such as SMPL-X, as on each extracted frame there is one person centered, thus the parameters of the body model can be directly estimated from the frame. However, this method is computationally expensive, especially in crowded scenes, thus inference time can become high. Moreover, as persons are extracted from the frame general, contextual information may be reduced or lost [184].

On the other hand, bottom-Up approaches first estimate all keypoints on the frame and then associate the detected body parts with individual people. This approach has the advantage of lower computational cost, in case only joint coordinates are estimated. However, if the goal is to estimate

the full body mesh it requires additional body mesh approximation modules. The main challenge of this approach is the grouping of the joints and handling occlusions [184].

4.2.2. RGB Monocular 3D MoCap

Monocular RGB MoCap is a challenging problem, as in this case the real 3D scenario is projected to 2D, meaning one dimension is lost. Therefore the inverse projection is not a one-to-one projection, the 3D pose extraction from 2D images can lead to pose ambiguities, as different 3D poses can be projected to similar 2D poses. As only one viewpoint is observed the target person suffers from self-occlusion and occlusions from other objects. Approaches to address the occlusions are presented in Sec. 4.3.1. RGB Monocular 3D MoCap can be organized into two main categories, skeleton-only and human mesh recovery approaches.

Skeleton only For skeleton-only approaches one straightforward method is to directly estimate the 3D skeleton parameters from the 2D images [150, 90, 121, 120].

As 2D HPE is a wildly explored area, utilizing that knowledge, by 2D to 3D lifting is a popular approach. In this case, a SOTA pre-trained 2D HPE network is used to estimate the 2D poses, and then utilizing this in the second stage the 3D parameters are inferred. These approaches usually outperform direct 3D skeleton estimation, as they build on the existing SOTA 2D HPE knowledge [138, 58, 160, 185, 151, 106, 103, 23, 81]. As the human body can be naturally represented as a graph, where joints representing the nodes and bones the edges, it is natural to apply Graph Convolutional Networks (GCNs), to address the 2D to 3D lifting problem. [183, 180, 94, 33, 34]

During skeleton-only approaches it is advantageous to utilize a kinematic model, as these can provide kinematic constraints for the pose estimation, thus realistic poses are estimated, which can improve performance by introducing prior knowledge. [187, 75, 80, 169, 157, 109, 108, 76, 45]

As MoCap is performed not only on one image, but commonly it is required to be performed on videos, where temporal consistency is essential, thus including this information into the MoCap pipeline can improve the performance and robustness of MoCap. This approach aims to mitigate the effect of short-term occlusions, by exploiting temporal information from the pose sequences [53, 188, 159, 152, 123, 29, 14, 36]. This can be achieved by including LSTMs [53], spatial-temporal relationships and constraints, such as bone length [14, 36, 87], or temporal convolution [123].

Human mesh recovery Human mesh recovery networks take advantage of sophisticated human body models. There are two main directions to estimate the model parameters. Firstly, optimization-based methods fit a parametric body model to 2D observations in an iterative manner, leading to accurate image model alignments, but are often slow and sensitive to the initialization [154, 148, 11, 124, 78, 71]. On the other hand, regression-based methods, that use a deep network to directly estimate the model parameters from pixels, tend to provide a reasonable, but not pixel-accurate result, while requiring huge amounts of supervision [71, 114, 122, 63]. A recent approach, FrankMocap, aims to improve the full-scale human body regression, including body, face, and hands, by utilizing separate SOTA expert modules for each of them, thus taking full advantage of recent high-performance modules and then eventually integrate them to one SMPL-X model with an integration module [131].

In order to close the gap between the two approaches SPIN (SMPL oPtimization IN the loop) [71] was proposed, combining together regression and optimization.

An adversarial training approach was proposed, "Video Inference for Body Pose and Shape Estimation" (VIBE) [69], to take advantage of several MoCap datasets combined together, called (AMASS) [100]. The adversarial training encourages the regressor to produce realistic and accurate motions and decide if the proposed motion capture is realistic. The architecture takes advantage of a temporal generation network trained together with a motion discriminator [69]. The idea of temporal consistency was further improved, by focusing on the past and future frames' temporal information without being dominated by the current static features [32].

One of the most recent architectures, which performs the best on the 3DPW dataset [102], utilizes transformer architectures [92], named Mesh Graphormer, which is an improved version of the Metro architecture [91]. Mesh Graphormer is a graph-convolution-reinforced transformer, that aims to combine the advantages of transformer-based approaches, such as modeling non-local interactions among 3D mesh vertices and body joints, and the advantage of graph-convolutions of exploiting neighborhood vertex interactions based on a pre-specified mesh topology [92].

4.3. Approaches to solving clinical challenges of 3D MoCap

4.3.1. Occlusions and viewpoints

One of the main challenges of 3D MoCap and clinical patient monitoring is the handling of occlusions. Therefore there are several approaches to addressing this issue.

First of all, it can be approached from a data acquisition aspect, such as acquiring multi-modal data, specially targeted for clinical in-bed monitoring with blanket occlusions, including RGB, LWIR, depth and pressure sensors [95, 174, 130]. Then fusing these modalities together can improve the performance of MoCap [95, 174, 130, 96]. Another approach was proposed utilizing synthetic data by simulating the blanket occlusion on depth data and train the architectures with this data [1]. Additionally, a synthetic blanket occlusion augmentation pipeline [16] was proposed and demonstrated on 3DPW [102], then the trained architecture was evaluated on a real-world clinical RGB-D-IR action recognition and semi-synchronized MoCap dataset [17].

Moreover, multi-view approaches are able to overcome some occlusions, as if one keypoint is occluded on one viewpoint, then from other viewpoints it can still be visible. It has a clear advantage when it is utilized in the multi-view scenario; however, these methods usually need large

memory and expensive computational cost, especially for multi-person 3D HPE, as several video streams have to be inferred. Most of these strategies utilize a multi-stage approach [60, 57, 42, 39, 38, 9, 129, 12, 8], where first the 2D poses are estimated for each frame, then the poses are clustered across views, from these clusters the 3D poses are reconstructed based on triangulation, and finally linked over time [9, 12]. The current top performing approach is TesseTract [128], which instead of formulating the learning in a multi-stage manner combines together person detection (3D CNN), tracking (4D CNNs) and pose estimation into one end-to-end network, with separate sub-losses for each subtask. During the person detection, the feature maps coming from all the camera views are aggregated into a common 3D voxelized volume, thus there are no assumptions posed on the available number of viewpoints, therefore it is able to operate in a monocular setting as well.

As discussed earlier in Sec. 4.2.2 temporal connections are essential for consistent MoCap in videos [105, 178, 28]. It is especially advantageous to address tracking with spatiotemporal representation learning, utilizing Spatio-Temporal Descriptors (4D volumes), as shown in [128]. This approach has a two-fold advantage, once, this temporal context allows to extrapolate/interpolate occluded joints and handle pose or appearance ambiguities, secondly, it improves tracking by generating a descriptor that overlaps with adjacent frames [128]. Another approach proposed, is temporal gated convolutions to recover missing poses and address the occlusion issues in the pose estimation, inspired by the image inpainting tasks [46]. As a post-processing step, refining the predicted trajectories through a Kalman filter in clinical environment [25], or learning motion priors, to smooth the movement [181] can be advantageous.

Occlusions were addressed in training time, with occlusion-aware training [30, 28, 134, 97]. During training time, occlusions were added as data-augmentation, either by occluding, practically set to zero, the whole frame, or keypoints, or some arbitrary area of the frame. The shape of the synthetic occlusions, also influenced performance, from the investigated basic shapes, the circle shape occlusion decreased the performance the most for architectures not trained specifically for occlusions [134]. Occlusion augmentation during training time, also had a regularizing effect, improving baseline performance [134].

Depending on the viewpoint, different self-occlusions are present in the 2D projection. In order to improve view invariance and occlusion robustness of the architecture metric learning approaches were proposed. It enables to capture measures of similarity between inputs, map close together similar 3D poses, and further away different 3D poses in the embedding space [97]. It is commonly achieved by utilizing contrastive loss [115, 113, 48, 26, 13] or triplet loss (based on tuple ranking) [166, 158, 135, 51]. These methods, train the networks with multiple inputs. In the case of triplet loss, a reference input (anchor) is compared to a matching input (positive sample) and a non-matching input (negative sample). The loss function is based on the distance between their embeddings, see Eq. 1,

$$L(A, P, N) = max(\| f(A) - f(P) \|^2 - \| f(A) - f(N) \|^2 + \alpha; 0) \quad (1)$$

where A-Anchor, P-Positive sample, N-Negative sample, α - margin, f- embedding. During training, the distance of the positive sample is minimized from the anchor and the distance of the negative sample is maximized.

The probabilistic re-formulation of this approach is described in detail in [97]. With this probabilistic approach input ambiguities of 2D poses originating from projection and occlusion can be mitigated. Deterministic embedding has to be close to either one of the similar embedding clusters or stay in the middle, even when there is uncertainty due to occlusion, or ambiguities in 2D pose [97]. On the other hand, probabilistic formulation, allows the embedding to hedge their bets across the embedding space [113], [97]. View invariance can be also improved, by estimating the camera parameters and perspective, and then utilizing these and the image features together to regress 3D body shape and pose as described in SPEC [70].

4.3.2. Low resolution 3D MoCap

Most of the 3D MoCap, HPE, and hand MoCap datasets are based on lab-captured HD data, which is not always available in the wild. From the clinical and monitoring perspective, storing and processing large amounts of data is not desirable. Moreover, during monitoring, it is necessary to provide precise 3D MoCap for hands and face as well. As the cameras have to be positioned relatively far away for monitoring, the extracted hands from the captured videos are only available in low resolution (LR).

There are a very limited number of approaches addressing 3D human pose and shape estimation on low-resolution videos [171, 170, 156]. However in low-resolution image processing, there are approaches to address this issue in 2D body pose estimation [107], face recognition [44, 31], image classification [163], image retrieval [149, 112], and object detection [50, 82]. There are two common approaches to handling low-resolution images. One of them is applying super-resolution or image enhancement techniques to the input [31, 163, 50] or in the feature space [44, 149, 112, 82], however it may result in unpleasant artifacts. The other approach is to simply train one model for each resolution, which is impractical in many realistic applications [170, 171].

In order to overcome the limitations of handling low resolution with the aforementioned techniques, RSC-Net was proposed, which consists of a Resolution-aware network, a Self-supervision loss, and a Contrastive learning scheme in [170], this approach was extended to videos and texture reconstruction in [171]. This approach enforces the feature and scale consistency, across different resolutions [170].

It is possible to utilize in the case of low-resolution videos super-resolution DL architectures, to improve the quality of the videos, thus decreasing the domain gap between pre-trained MoCap datasets and architectures, thus

improving results [139, 133, 85, 22]. There are two main categories of super-resolution (SR) techniques, SR on images, and SR on videos [21, 20, 88, 59]. There are extensive reviews on the topic, for details see survey papers [86, 24, 164], also reviewing previous reviews. However, SR techniques on synthetic data usually overestimate the capacity to super-resolve real-world images and domain-specific images [24]. The most successful modern SR architectures are based on generative adversarial networks (GANs) [136, 79, 20], auto-encoder [116, 146] and transformer architectures [89, 88]. SOTA video SR techniques aim to make full use of complementary information across frames to reconstruct the high-resolution sequence [88, 59].

4.3.3. Others

As essentially a post-processing step of 3D MoCap, physics-based human motion estimation, and synthesis were proposed, that corrects imperfect image-based pose estimations by enforcing physics constraints and reasons about contacts in a differentiable way [168]. This can be also achieved by including an inverse kinematics solver, which can significantly improve occlusion robustness [127].

To further improve results and improve domain transfer instance aware re-colorization might be applied [74, 145]. Additionally, video inpainting techniques, such as [177, 186], might be considered to be utilized as a pre-processing step to remove some of the occlusions. However, the performance impact of these techniques has not been evaluated for downstream tasks such as 3D MoCap or action recognition. However, addressing these challenges are beyond the scope of this review.

5. Semiology-based automated seizure classification

One of the most challenging applications of clinical in-bed action recognition is epileptic seizure classification based on seizure semiology in EMUs. Which includes all of the clinical challenges discussed above. Therefore in the following a critical review of this sub-field is presented, which well represents the general principles, remaining challenges, and possible solutions for clinical in-bed action recognition applications.

The motivation to quantitatively analyze seizure semiology, specific movements during the seizures, and classify epileptic seizures, was expressed by several earlier research [35, 64, 2, 101, 3, 4, 5, 7, 54, 126, 55, 6, 125].

These approaches utilize classical computer vision based [35], image or video classification [2, 3, 101, 5, 4, 126, 55], keypoint, or keypoint stream [7, 6, 54] and action recognition based approaches [64, 66, 125], a summary is presented from these approaches in Table 2. All of these are utilizing private clinical datasets that are not publicly accessible due to clinical and personal data protection concerns. The lack of a standard public benchmark dataset makes it hard to evaluate and directly compare these methods.

These approaches highlight the importance of spatio-temporal information of the movements during the seizures and utilizing an action recognition approach [64, 66, 125]. Moreover, to guide the architecture training and relevant feature extraction a multistep approach is beneficial. In this context, the initial phase of extracting keypoint features plays a crucial role in isolating the most significant movement features, which are then utilized in the subsequent phase of classification [7, 6, 54, 125]. These show that a 3D MoCap based first step to extract these features describing the movement, as precisely as possible, during the full seizure would be essential.

However, the above mentioned approaches do not utilize 3D MoCap, a spatio-temporally consistent detection, tracking, and 3D HPE. As the 3D keypoint feature extraction is carried out frame by frame it neglects temporal connections, which provide a noisy input for the classification step [54]. Therefore exploring the possibility to approach the first step with 3D MoCap, to improve the spatio-temporal consistency of the extracted features in a clinical environment is important.

Moreover, several works reported very promising performance on small datasets utilizing the same patient both in training and testing, validation sets [5, 126, 7, 6, 125] or highly aggregated performance metrics, such as classifying snippets with low performance, but aggregated up to the patient level to achieve very high-performance [4, 125].

The issue of utilizing the video data from the same subject in train and test sets is well known. Subject-specific features can cause data leaks, which may misleadingly improve performance, but not due to the target basis of classification. Such as utilizing the raw images, and videos with facial recognition-based feature extraction is very concerning that subject-specific features were contributing more to the classification and not seizure-type specific features. It is suggested by the low Leave-One-Subject-Out (LOSO) Cross Validation (CV) performance in many works. This holds for the keypoint feature extraction-based approaches, as this step can be still a subject-specific representation of the seizure. Moreover, the extracted keypoints represent the distance in space of 3D joints, which represents the body size and proportions, which is a highly subject-specific feature and may lead to overfitting to this subject-specific feature. Certainly, another explanation for the low LOSO-CV performance may be that there could be seizure-specific features not represented in the training set, due to the small available dataset. Nevertheless, separating the patients to have all the seizures of a patient in either the training or the testing sets is essential to undoubtedly prove the generalizability of the proposed system.

The limitation of the reported highly aggregated results is that a small variation of the data can cause this aggregated performance to drop significantly, as it relies on sub-results with significantly lower performance, which raises questions about the generalization of the network. Moreover, an approach, which aggregates the results through the full seizure may not be utilized well for near real-time applications.

Table 2: Automatic semiology-based deep learning-based seizure classification publications in this domain. (ETLE - Extra Temporal Lobe Epilepsy, MTLE - Mesial Temporal Lobe Epilepsy, FOS - focal onset seizures, TCS - focal to bilateral tonic-clonic seizures, std – standard deviation, FC - Fully Connected, LSTM - Long Short Term Memory, HAR - Human Action Recognition, AGCN - Adaptive Graph Convolutional Network)

Cat.	Author	Classes	Feature extraction	Classification Method	Performance (Cross subject, sequence wise)	Notes
Class. CV	Cunha et al. [35]	TLE ETE	Semi-automatic tracking	Optical flow + depth	*Jerk quantification* mean=345 ms^{-3} vs 172 ms^{-3}, p=0.05	Discriminated both seizure MOI groups with statistically significant levels, NeuroKinect Quantification of movements
Image or video classification	Achilles et al. [2]	Seizure No seizure	CNN from IR-D images	Image classification	AUC: 0.78	Single frame approach (posture recognition)
	Ahmedt-Aristizabal et al. [3]	MTLE ETLE	Face, Hand tracking, 3D pose estimation	Face,Hand: CNN+LSTM, change of pose: LSTM	Average accuracy:56.31% (best, just body)	Face body and hand inputs, very high std, lack of action recognition considerations
	Maia et al. [101]	TLE ETLE	InceptionV3 (Image)	FC	AUC 0.65	Probably overfits
	Ahmedt-Aristizabal et al. [5]	MTLE ETLE	Facial landmark metric and Facial features	SVM and CNN-LSTM	Average accuracy: 50.85%	Subject specific accuracy 95.19%, but susceptible to overfit to subject specific facial features
	Ahmedt-Aristizabal et al. [4]	MTLE ETLE	CNN-LSTM	Cosine similarity with Seizure feature MoCap library	Average accuracy: 66.48%; 62.19%	Promising aggregated cosine similarity results through all seizures of each patient, AUC: 0.9703
	Pothula et al. [126]	Gelastic Dacrystic No seizure	Facial emotion recognition (5 class)	Random Forest (estimators=60)	Acc. *(not CrossSubj.)*: 98.8%	Only 4 patients, same videos in train and test set, oversampling in dataset generation is concerning
	Hou et al. [55]	ES PNES	Transformer	FC	f1-score: 0.82	Pre-trained transformer on clinical videos, with ImageNet pre-trained ResNet-152 CNN backbone
Keypoint, keypoint stream	Ahmedt-Aristizabal et al. [7]	MTLE ETLE	Facial features and Upper limb 3D keypoints	Fusion of extracted features, LSTM	Average accuracy: 58.49%	Subject specific accuracy 92.10%, but susceptible to overfit to subject specific facial features and posture coordinates
	Ahmedt-Aristizabal et al. [6]	MTLE ETLE	3D Face reconstruction, mouth ROI	LSTM-FC	Average accuracy: 69.8%	Subject specific accuracy 89%,but susceptible to overfit to subject specific facial features
	Hou et al. [54]	ES PNES	3D Human pose keypoint, facial landmark stream	AGCN	f1-score: 0.76	Knowledge distillation from video stream for training the AGCN
Action recognition	Karácsony et al. [64]	TLE FLE	I3D action recognition network	LSTM-FC	f1-score: 0.844+-0.042 (AUC: 0.90+-0.04)	End-to-end action recognition
	Karácsony et al. [66]	TLE FLE No Seizure	I3D action recognition network	LSTM-FC	f1-score: 2 class: 0.833±0.061 3 class: 0.763±0.083	End-to-end action recognition
	Perez et al. [125]	FOS TCS	Sampled snippet level pre-trained action recognition network	bidirectional LSTM-FC	f1-score *(not CrossSubj.)*: 0.987	HAR, seizure level aggregation, same patient in train and test split is concerning

6. Discussion

6.1. Clinical in-bed monocular video-based action recognition

Clinical in-bed video-based action recognition can be utilized in many diagnosis support scenarios, such as epileptic seizure and sleep disease monitoring, classification, and real-time alarms, however still face many computer vision challenges (Fig. 1), which are rarely and only partially addressed in the literature. Here, the perspective of the epileptic seizure classification was reviewed, which represents one of the most challenging clinical in-bed action recognition scenarios, with spatio-temporal features, several occlusions, uncommon viewpoint, and many other challenges. The current state-of-the-art solutions for this clinical challenge are further discussed in section 6.4. From the clinical perspective, movement quantification is essential in clinical diagnosis, therefore it is advantageous to include it already in these systems.

As clinical data is scarce, it is essential to utilize the labeled clinical action data efficiently. Collection and labeling of clinical video data, for example, epileptic seizure data, can spawn a very long time, as collection depends on the capacity of hospitals, patients manifesting the seizures, and ground truth labeling is only available after full clinical diagnosis. Therefore it is advantageous to utilize already existing data; however, this legacy data may be low-resolution RGB or IR videos with varying quality of the videos. For these differences, the designed action recognition approach has to be adopted. Moreover, as the data collection takes place over an extended period of time, it has to be considered in the implemented MLOps approach as well. Training action recognition networks end-to-end, mapping the raw video data to the action classes, is computationally expensive. On the other hand, a two-stage approach with MoCap as the first stage can alleviate this challenge. Moreover, the MoCap architecture can heavily rely on transfer learning to tackle the clinical challenges, as further discussed in Sec. 6.3.

A two-stage action recognition approach, where movement features are extracted with a 3D MoCap network, is advantageous. This way training for detection, tracking, and extraction of motion features is clearly separated from the clinical features of the movement, thus several other datasets can be utilized in a transfer learning approach to training the first stage. On the contrary, in end-to-end approaches, the clinical and general features are not separated clearly, thus during training of the network clinical data may contribute to the movement feature extraction training as well, which may lower clinical performance and generalization of the network.

This approach however has some drawbacks. The action recognition heavily relies on the performance of the MoCap approach, therefore the noise and errors generated by the MoCap can propagate to the classification, thus can reduce performance. Therefore it is essential to enforce spatio-temporal stability as much as possible for the MoCap system.

Here we have to note that the near real-time performance of either of the presented architectures vary, largely dependent on available computational resources and code optimization, such as quantization and pruning, further influenced by the software framework and the efficiency of the implementation. While some models may run efficiently on mid-range gaming PCs, others require high-end GPU clusters. Although the inference phase of these architectures is significantly less demanding than training, its resource intensity still varies based on model complexity and input resolution. Additionally, most of these methods are designed to process sequences of frames rather than just individual ones, enabling them to capture essential temporo-spatial features. However, these approaches may introduce an inherent lag, potentially spanning a few seconds, depending on the specific implementation. Therefore, while these systems can achieve near real-time performance, most of them do not strictly conform to real-time processing standards. In certain contexts, such as in-bed monitoring applications, this level of performance may be considered effectively real-time, despite the slight delay in processing.

In summary for clinical in-bed monocular video-based action recognition a 2 stage approach is promising with Motion capture as a first stage, as it can provide clinically relevant quantification of movements, contribute to efficient utilization of clinical data and establish a practical MLOps approach.

6.2. Clinical monocular video based 3D MoCap for in-bed monitoring

As discussed earlier there are several challenges to applying 3D MoCap for in-bed scenarios (Fig. 1); however, for action recognition for in-bed scenarios, it is essential to address these. It has to be emphasized that the goal is to extract spatio-temporal features of the movement from the raw video (3D MoCap), not just spatial features from images, as in some human pose estimation approaches. Utilizing temporal connections additional to the spatial features is essential for downstream tasks and may improve spatio-temporal stability of this feature extraction. Moreover, utilizing 3D MoCap as the first stage for action recognition introduces many advantages over an end-to-end action recognition approach, such as inherent quantitative movement parameters of the action, which may provide an explanation of movement classes; additionally, it introduces further datasets and options to apply transfer learning (See Sec. 6.3).

In the following, we discuss the most promising directions for these clinical challenges.

6.2.1. Occlusions and viewpoints

One of the main challenges of in-bed MoCap is the occlusions present. This can be divided into three main categories: self-occlusions, attending clinicians, and blanket occlusion.

Self occlusions are generally addressed by 2D and 3D HPE literature, such as multi-view approaches, utilizing

triplet loss and metric learning approaches, which already address partially the viewpoint challenges as well.

Attending clinicians are present on the scene in a relatively short temporal range, moving around, causing a dynamic occlusion. However, this occlusion is present usually when important data is recorded, such as seizures, therefore it is an important aspect to address. Due to the dynamic nature of the occlusion, (i.e they move around), it can be partially addressed with temporal feature considerations included in the MoCap. The attending clinicians close to the patient raise the issue of keypoint assignment and grouping, thus top-down approaches might be more advantageous to utilize, where the full person is identified first and then cropped. Moreover, in this case, the higher performance requirement of such approaches for crowded scenes is negligible, as only a few people are in the scene.

The most challenging category is blanket occlusion as it is prevalent almost the entire monitoring, covering a significant portion of the target for extended periods of time. There are attempts to solve this challenge of clinical in-bed HPE. Occlusion-aware training already showed increased performance with occlusion generated just as black hold-out areas, thus utilizing real blanket occlusions, which still have useful features is advantageous. Datasets addressing this challenge are rare, one containing only multimodal images of in-bed poses, with high occlusion (SLP [96, 95]), and ground truth joint coordinate locations. However, as it only contains images it neglects temporal features, which would be essential to extract more information about limbs moving under the blanket to improve MoCap performance. Acquiring the ground truth joint positions under blanket occlusion is not straightforward, such as when utilizing long wave infrared (LWIR), as in SLP dataset it is not possible to track quick movements. IMUs could be utilized on each joint; however, it would require attaching an IMU to each point of interest, which already significantly alters the visual representation of the scene, moreover, these can be dislodged due to friction with the blanket. As the acquisition of new in-bed data with blanket occlusion and ground truth labels of the joint coordinates is so challenging it is highly beneficial to utilize pre-existing datasets with synthetic augmentation [16, 17]. Synthetic augmentation of already existing large MoCap datasets provides the opportunity to generate virtually unlimited augmented data, with varying blankets in color, thickness, position, and other parameters. Although there is a domain shift caused by synthetic data, the opportunity to generate massive amounts of data with ground truth labels outweighs the disadvantages.

In order to address the viewpoint challenge it is essential to utilize large datasets well covering all the viewpoints; however, existing datasets do not fully represent these, especially lacking the common viewpoint in hospital monitoring (Fig. 1). Thus it would be advantageous to develop or extend a dataset that represents these viewpoints as well. Furthermore, it is advantageous to utilize multi-view data for training. Either by utilizing metric learning to map 2D poses from different viewpoints of the same 3D pose close to each other in feature space or achieving it by training end-to-end with multi-view input, such as TesseTract [128], which inherently has these advantages. To further improve viewpoint invariance of the estimated pose it can be advantageous to estimate the camera parameters (SPEC [70]).

In summary, these challenges do not require clinically relevant data, containing clinical actions such as seizures, most of these can be addressed based on already publicly available MoCap and HPE datasets. In the case of general occlusions spatio-temporal descriptors can assist to interpolate or extrapolate the occluded keypoints, moreover utilizing synthetic dataset augmentation can be effective to address blanket occlusions, as a specific case of occlusion-aware training. Besides training on large datasets containing many viewpoints, utilizing multiview or multimodal datasets is advantageous, additionally, including metric learning, and triplet loss in the training strategy is highly desirable.

6.2.2. Low resolution

The challenge of dealing with low-resolution data is the most prevalent for legacy data, however as data collection and labeling in the clinical domain require a long time span it is essential to address to broaden the amount of available clinical data. Moreover, it is relevant for more detailed motion capture including the hands, fingers, and facial expressions as well, due to the common viewpoint in clinics monitoring the full body from a distance, which will result in relatively low-resolution segments representing these even on 1080p videos. For example, most hand MoCap systems are optimized for close-up monitoring of hands, utilizing HD data.

Training networks for multiple resolutions, including low-resolution videos is not practical. Super-resolution techniques, and image enhancement techniques to the input or in the feature space are commonly utilized techniques in 2D HPE; however, may introduce unwanted artifacts. Therefore resolution-aware networks with a contrastive learning scheme may be the most beneficial to address this challenge, such as RSC-Net [171, 170].

6.3. Datasets and Transfer learning for clinical MoCap and action recognition

As discussed earlier, labeled clinical data is very scarce, moreover, it is very challenging to acquire additional data, therefore efficient data utilization is essential. Namely utilizing the clinical data only for clinical feature classification. As presented in Sec. 5 approaches heavily relying on transfer learning to extract low-level image, video, movement or action features perform significantly better, as they are able to incorporate a lot more training data.

In conclusion, we can articulate the requirement deduced from previous works to propose a 2 stage approach to address clinical in-bed action recognition, utilizing transfer learning for feature extraction such as 3D MoCap and action recognition.

In order to achieve efficient data utilization, exploiting several levels of transfer learning the following options may

Table 3
Optional levels of transfer learning to be considered for clinical action recognition

Levels of transfer learning	Feature extraction	1. stage	2. stage	Clinical data utilization
1. Image/object recognition, segmentation	spatial	yes	no	no
2.a Image-based Human Pose estimation 2D-3D, human mesh recovery	spatial	yes	no	no
2.b Image-based Face, hand features and keypoint estimation	spatial	yes	no	no
3.Video-based: Motion capture: detection, tracking, and HPE, hand, face tracking	Spatio-temporal	yes	no	no
3.b + challenges - occlusion aware training, low resolution, etc.	Spatio-temporal	yes	no	no
4. Skeleton based action recognition	Spatio-temporal features -> classification	no	yes	no
4.b Emotion recognition	Spatio-temporal facial features -> classification	no	yes	no
5. Clinical action recognition e.g. Semiology and seizure classification	Spatio-temporal features -> classification	no	yes	yes

be utilized, which are summarized in Tab. 3. The first level of transfer learning is sourcing training data from image/object recognition and segmentation datasets, which can be used for pre-training the architecture such as the authors did in I3D[19]. The second level is utilizing data from *2.a* Image-based Human Pose estimation 2D-3D, human mesh recovery, including *2.b* face, hand features and keypoint estimation. As there are HPE datasets that contain only images, thus only the spatial features of the poses can be extracted, these can be datasets such as SLP a specialized in-bed occlusion dataset [95] or face recognition datasets The third level, *3.a* consists of the MoCap, the final goal of the first stage, which extracts the spatio-temporal features of movements, and other relevant keypoints on the hands, and face. Here *3.b* the additional challenges can be addressed as well, such as occlusions, by occlusion-aware training. The fourth level, *4.a and 4.b*, are the skeleton-based action recognition and emotion recognition to extract features from the movements extracted by the first stage. Eventually, the clinical data has to be utilized only in the fifth stage for the transfer learning of the action recognition networks for clinical purposes. In this stage, the transfer learning from action recognition level 4 is essential to train a generalizing network and not to fit for patient-specific clinical features.

Utilizing a two-stage network is advantageous from an MLops approach, as the method of motion capture of the skeleton can be replaced by improved architectures, may include additional modalities, or even rely on a different data capture method. This way the second stage, the clinical action classification, which only relies on the tracked keypoints does not require re-training from scratch only fine training, which promotes the flexibility and rapid development of the architecture, contrary to the end-to-end action recognition approaches.

The drawback of such as 2-stage architecture is the propagation of error, from the first stage to the second stage, which is highly dependent on the quality of the spatio-temporal feature extraction of the first stage.

6.4. Current state of semiology-based seizure classification

Many earlier approaches neglect the seizure semiologys' spatio-temporality of features, more recent approaches utilizing some way of spatio-temporal feature extraction and classification achieved higher performance.

In the literature, a common challenge is the scarcity of clinical data, thus efficient data utilization is essential for future approaches. Therefore, clinical data should only be utilized for the action classification highly depending on transfer learning. The transfer learning can rely on many levels as discussed above, the most advantageous approach is to separate the motion capture, essentially the spatio-temporal feature extraction of movements and the action classification, and utilize clinical data, such as videos from epileptic seizures, only for fine training of the action classification architecture.

One major barrier to the existence of a publicly available benchmark clinical datasets for comparing different approaches stems from concerns over clinical and personal data protection. Patients can potentially be identified from raw video data. However, adopting a two-stage approach, where the first stage involves extracting skeletons from the raw videos, could pave the way for creating a shareable public benchmark dataset. These extracted skeletons, being privacy-preserving, can be exchanged among clinics and research groups and made publicly available. This strategy has the potential to mitigate the current scarcity of clinical data for research purposes.

A good practice is to separate the action videos by patients, so the videos from one patient are only included in either train or validation or test set, thus preventing any data leakage of patient-specific features, thus ensuring generalization of the architecture. In some articles, this data

separation was not considered, which might have produced misleadingly high performances.

The explainability of classification would be an important aspect to provide for the eventual users of such diagnostic support systems, however, this issue has not been addressed yet in the analyzed articles.

7. Future challenges and research directions

In the future, the main challenge is related to stage 1, developing a spatio-temporally stable and robust MoCap, including body, hands, and face, to extract the movement features and the challenges arising from this. These challenges include occlusion handling, with a special focus on blanket occlusions, the low resolution of the full scenes of legacy data, or the relatively low resolution of the face and hand, addressing the bias arising from uncommon viewpoints and extending the architectures to be able to handle IR videos to be able to monitor 24/7.

In stage 2, the clinical action recognition, an important aspect of future research would be to include the explainability of the classifications, which is essential in diagnostic support systems

8. Conclusion

In conclusion, monocular video-based clinical in-bed action recognition was reviewed from an epileptic seizure classification perspective. The main challenges of such systems were identified and possible approaches to address these were reviewed. The feasibility of semiology-based automated seizure classification was already established by several papers reviewed here, from these it can be deduced that a 2-stage quantified approach with motion capture and action recognition is the most promising research direction in the future.

Clinical action recognition is required for several diagnosis support scenarios, such as seizure diagnosis support, sleep monitoring, and clinical research among others, therefore advancing such systems can have a significant impact in many fields.

9. Acknowledgements

We thank members of the Epilepsy Center, Department of Neurology, University of Munich (LMU), Munich, Germany, and the Neurophysiology Unit, Neurology Department, Centro Hospitalar Universitário de São João, E.P.E. (CHUSJ-UPorto), Porto, Portugal, for providing clinical requirements for this work along the years of fruitful collaboration, namely Nicholas Fearns, Dr. Anna Mira Loesch Biffar, PD Dr. Dr. Christian Vollmar, Prof. Dr. med. Jan Rémi, Prof. Dr. Soheyl Noachtar (LMU) and Dr. Ricardo Rego (CHUSJ-UPorto).

This work was partially funded by Fundação para a Ciência e a Tecnologia under the scope of the CMU Portugal program (Ref PRT/BD/152202/2021).

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020. DOI 10.54499/LA/P/0063/2020 |
https://doi.org/10.54499/LA/P/0063/2020

CRediT authorship contribution statement

Tamás Karácsony: Conceptualization of this study, Methodology, Writing the Original Draft, Review & Editing and Visualization. **László Attila Jeni:** Conceptualization of this study, Methodology, Review & Editing. **Fernando De la Torre:** Conceptualization of this study, Methodology, Review & Editing. **João Paulo Silva Cunha:** Conceptualization of this study, Methodology, Review & Editing, Supervision.

References

[1] Achilles, F., Ichim, A.E., Coskun, H., Tombari, F., Noachtar, S., Navab, N., 2016a. Patient MoCap: Human pose estimation under blanket occlusion for hospital monitoring applications, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag. pp. 491–499. doi:10.1007/978-3-319-46720-7_57.

[2] Achilles, F., Tombari, F., Belagiannis, V., Loesch, A.M., Noachtar, S., Navab, N., 2016b. Convolutional neural networks for real-time epileptic seizure detection. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 6, 264–269. doi:10.1080/21681163.2016.1141062.

[3] Ahmedt-Aristizabal, D., Fookes, C., Denman, S., Nguyen, K., Fernando, T., Sridharan, S., Dionisio, S., 2018a. A hierarchical multimodal system for motion analysis in patients with˜epilepsy. Epilepsy & Behavior 87, 46–58. doi:10.1016/j.yebeh.2018.07.028.

[4] Ahmedt-Aristizabal, D., Fookes, C., Denman, S., Nguyen, K., Sridharan, S., Dionisio, S., 2019a. Aberrant epileptic seizure identification: A computer vision perspective. Seizure 65, 65–71. doi:10.1016/j.seizure.2018.12.017.

[5] Ahmedt-Aristizabal, D., Fookes, C., Nguyen, K., Denman, S., Sridharan, S., Dionisio, S., 2018b. Deep facial analysis: A new phase I epilepsy evaluation using computer vision. Epilepsy & Behavior 82, 17–24. doi:10.1016/j.yebeh.2018.02.010.

[6] Ahmedt-Aristizabal, D., Nguyen, K., Denman, S., Sarfraz, M.S., Sridharan, S., Dionisio, S., Fookes, C., 2019b. Vision-Based Mouth Motion Analysis in Epilepsy: A 3D Perspective, in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, Institute of Electrical and Electronics Engineers Inc.. pp. 1625–1629. URL: https://ieeexplore.ieee.org/document/8857656/, doi:10.1109/EMBC.2019.8857656.

[7] Ahmedt-Aristizabal, D., Nguyen, K., Denman, S., Sridharan, S., Dionisio, S., Fookes, C., 2018c. Deep Motion Analysis for Epileptic Seizure Classification, in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, Institute of Electrical and Electronics Engineers Inc.. pp. 3578–3581. doi:10.1109/EMBC.2018.8513031.

[8] Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S., 2016. 3D Pictorial Structures Revisited: Multiple Human Pose Estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 38, 1929–1942. doi:10.1109/TPAMI.2015.2509986.

[9] Belagiannis, V., Wang, X., Schiele, B., Fua, P., Ilic, S., Navab, N., 2015. Multiple human pose estimation with temporally consistent 3D pictorial structures, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer, Cham.

pp. 742–754. URL: https://link.springer.com/chapter/10.1007/978-3-319-16178-5_52, doi:10.1007/978-3-319-16178-5_52.

[10] Blum, D.E., Eskola, J., Bortz, J.J., Fisher, R.S., 1996. Patient awareness of seizures. Neurology 47, 260–264. doi:10.1212/wnl.47.1.260.

[11] Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J., 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 561–578. doi:10.1007/978-3-319-46454-1_34.

[12] Bridgeman, L., Volino, M., Guillemaut, J.Y., Hilton, A., 2019. Multi-person 3D pose estimation and tracking in sports, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 2487–2496. doi:10.1109/CVPRW.2019.00304.

[13] BROMLEY, J., BENTZ, J.W., BOTTOU, L., GUYON, I., LECUN, Y., MOORE, C., SÄCKINGER, E., SHAH, R., 1993. SIGNATURE VERIFICATION USING A "SIAMESE" TIME DELAY NEURAL NETWORK. International Journal of Pattern Recognition and Artificial Intelligence 07, 669–688. doi:10.1142/s0218001493000339.

[14] Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.J., Yuan, J., Thalmann, N.M., 2019. Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2272–2281. doi:10.1109/ICCV.2019.00236.

[15] do Carmo Vilas-Boas, M., Cunha, J.P.S., 2016. Movement Quantification in Neurological Diseases: Methods and Applications. IEEE Reviews in Biomedical Engineering 9, 15–31. doi:10.1109/rbme.2016.2543683.

[16] Carmona, J., Karácsony, T., Cunha, J.P.S., 2022a. BlanketGen – A synthetic blanket occlusion augmentation pipeline for MoCap datasets. arxiv:2210.12035 URL: https://arxiv.org/abs/2210.12035v1, doi:10.48550/arxiv.2210.12035.

[17] Carmona, J., Karácsony, T., Cunha, J.P.S., 2022b. BlanketSet – A clinical real word action recognition and qualitative semi-synchronised MoCap dataset. arXiv:2210.03600 URL: https://arxiv.org/abs/2210.03600v1http://arxiv.org/abs/2210.03600, doi:10.48550/arxiv.2210.03600.

[18] Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., Zisserman, A., 2018. A Short Note about Kinetics-600. arXiv:1808.01340v1 URL: http://activity-net.org/challenges/2018/evaluation.htmlhttp://arxiv.org/abs/1808.01340.

[19] Carreira, J., Zisserman, A., 2017. Quo Vadis, action recognition? A new model and the kinetics dataset, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 4724–4733. doi:10.1109/CVPR.2017.502.

[20] Chadha, A., Britto, J., Roja, M.M., 2020. iSeeBetter: Spatio-temporal video super-resolution using recurrent generative back-projection networks. Computational Visual Media 6, 307–317. URL: https://iseebetter.amanchadha.com., doi:10.1007/s41095-020-0175-7.

[21] Chan, K.C., Zhou, S., Xu, X., Loy, C.C., 2022. Investigating tradeoffs in real-world video super-resolution, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5952–5961. doi:10.1109/CVPR52688.2022.00587.

[22] Chan, K.C.K., Wang, X., Yu, K., Dong, C., Loy, C.C., 2021. BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond. IEEE Computer Society Conference on Computer Vision and Pattern Recognition .

[23] Chen, C.H., Ramanan, D., 2017. 3D human pose estimation = 2D pose estimation + matching, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 5759–5767. doi:10.1109/CVPR.2017.610.

[24] Chen, H., He, X., Qing, L., Wu, Y., Ren, C., Sheriff, R.E., Zhu, C., 2022. Real-world single image super-resolution: A brief review. doi:10.1016/j.inffus.2021.09.005.

[25] Chen, K., Gabriel, P., Alasfour, A., Gong, C., Doyle, W.K., Devinsky, O., Friedman, D., Dugan, P., Melloni, L., Thesen, T., Gonda, D., Sattar, S., Wang, S., Gilja, V., 2018. Patient-specific pose estimation in clinical environments. IEEE Journal of Translational Engineering in Health and Medicine 6. doi:10.1109/JTEHM.2018.2875464.

[26] Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations, in: 37th International Conference on Machine Learning, ICML 2020, pp. 1575–1585. URL: https://github.com/google-research/simclr.

[27] Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W., 2021. Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 13339–13348. URL: https://github.com/Uason-Chen/CTR-GCN., doi:10.1109/ICCV48922.2021.01311.

[28] Cheng, Y., Yang, B., Wang, B., Tan, R.T., 2020. 3D human pose estimation using spatio-temporal networks with explicit occlusion training, in: AAAI 2020 - 34th AAAI Conference on Artificial Intelligence, AAAI press. pp. 10631–10638. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6689, doi:10.1609/aaai.v34i07.6689.

[29] Cheng, Y., Yang, B., Wang, B., Wending, Y., Tan, R., 2019a. Occlusion-aware networks for 3D human pose estimation in video, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 723–732. doi:10.1109/ICCV.2019.00081.

[30] Cheng, Y., Yang, B., Wang, B., Wending, Y., Tan, R., 2019b. Occlusion-aware networks for 3D human pose estimation in video, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 723–732. doi:10.1109/ICCV.2019.00081.

[31] Cheng, Z., Zhu, X., Gong, S., 2018. Low-Resolution Face Recognition. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11363 LNCS, 605–621. URL: https://arxiv.org/abs/1811.08965v2, doi:10.1007/978-3-030-20893-6_38.

[32] Choi, H., Moon, G., Chang, J.Y., Lee, K.M., 2021. Beyond Static Features for Temporally Consistent 3D Human Pose and Shape from a Video, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1964–1973. URL: https://github.com/hongsukchoi/TCMR_RELEASE, doi:10.1109/CVPR46437.2021.00200.

[33] Choi, H., Moon, G., Lee, K.M., 2020. Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 769–787. URL: https://github.com/hongsukchoi/Pose2Mesh_RELEASE, doi:10.1007/978-3-030-58571-6_45.

[34] Ci, H., Wang, C., Ma, X., Wang, Y., 2019. Optimizing network structure for 3D human pose estimation, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2262–2271. doi:10.1109/ICCV.2019.00235.

[35] Cunha, J.P.S., Choupina, H.M.P., Rocha, A.P., Fernandes, J.M., Achilles, F., Loesch, A.M., Vollmar, C., Hartl, E., Noachtar, S., 2016. NeuroKinect: A Novel Low-Cost 3Dvideo-EEG System for Epileptic Seizure Motion Quantification. PLOS ONE 11, e0145669. doi:10.1371/journal.pone.0145669.

[36] Dabral, R., Mundhada, A., Kusupati, U., Afaque, S., Sharma, A., Jain, A., 2018. Learning 3D human pose from structure and motion, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 679–696. doi:10.1007/978-3-030-01240-3_41.

[37] Desmarais, Y., Mottet, D., Slangen, P., Montesinos, P., 2021. A review of 3D human pose estimation algorithms for markerless motion capture. Computer Vision and Image Understanding 212, 103275. doi:10.1016/J.CVIU.2021.103275.

[38] Dong, J., Fang, Q., Jiang, W., Yang, Y., Huang, Q., Bao, H., Zhou, X., 2021. Fast and Robust Multi-Person 3D Pose Estimation and Tracking from Multiple Views. IEEE Transactions on Pattern Analysis and Machine Intelligence URL: https://zju3dv.github.io/, doi:10.1109/TPAMI.2021.3098052.

[39] Dong, J., Jiang, W., Huang, Q., Bao, H., Zhou, X., 2019. Fast and Robust Multi-Person 3D Pose Estimation from Multiple Views. CVPR2019 URL: https://zju3dv.github.io/.

[40] Duan, H., Wang, J., Chen, K., Lin, D., 2022a. PYSKL: Towards Good Practices for Skeleton Action Recognition. arXiv:2205.09443 URL: https://arxiv.org/abs/2205.09443v1http://arxiv.org/abs/2205.09443, doi:10.48550/arxiv.2205.09443.

[41] Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B., 2022b. Revisiting Skeleton-based Action Recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2969–2978URL: https://github.com/kennymckormick/pyskl.

[42] Ershadi-Nasab, S., Noury, E., Kasaei, S., Sanaei, E., 2018. Multiple human 3D pose estimation from multiview images. Multimedia Tools and Applications 77, 15573–15601. URL: https://dl.acm.org/doi/abs/10.1007/s11042-017-5133-8, doi:10.1007/s11042-017-5133-8.

[43] Fürbass, F., Ossenblok, P., Hartmann, M., Perko, H., Skupch, A., Lindinger, G., Elezi, L., Pataraia, E., Colon, A., Baumgartner, C., Kluge, T., 2015. Prospective multi-center study of an automatic online seizure detection system for epilepsy monitoring units. Clinical Neurophysiology 126, 1124–1131. doi:10.1016/j.clinph.2014.09.023.

[44] Ge, S., Zhao, S., Li, C., Li, J., 2019. Low-Resolution Face Recognition in the Wild via Selective Knowledge Distillation. IEEE Transactions on Image Processing 28, 2051–2062. doi:10.1109/TIP.2018.2883743.

[45] Georgakis, G., Li, R., Karanam, S., Chen, T., Košecká, J., Wu, Z., 2020. Hierarchical Kinematic Human Mesh Recovery, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 768–784. doi:10.1007/978-3-030-58520-4_45.

[46] Gu, R., Wang, G., Hwang, J.N., 2020. Exploring severe occlusion: Multi-person 3D pose estimation with gated convolution, in: Proceedings - International Conference on Pattern Recognition, pp. 8243–8250. doi:10.1109/ICPR48806.2021.9412107.

[47] Guan, S., Xu, J., Wang, Y., Ni, B., Yang, X., 2021. Bilevel Online Adaptation for Out-of-Domain Human Mesh Reconstruction, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 10467–10476. doi:10.1109/CVPR46437.2021.01033.

[48] Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1735–1742. doi:10.1109/CVPR.2006.100.

[49] Han, F., Reily, B., Hoff, W., Zhang, H., 2017. Space-time representation of people based on 3D skeletal data: A review. Computer Vision and Image Understanding 158, 85–105. doi:10.1016/j.cviu.2017.01.011.

[50] Haris, M., Shakhnarovich, G., Ukita, N., 2021. Task-Driven Super Resolution: Object Detection in Low-Resolution Images, in: Communications in Computer and Information Science, pp. 387–395. doi:10.1007/978-3-030-92307-5_45.

[51] Hermans, A., Beyer, L., Leibe, B., 2017. In Defense of the Triplet Loss for Person Re-Identification. arXiv preprint arXiv:1703.07737 URL: http://arxiv.org/abs/1703.07737.

[52] Hoppe, C., Poepel, A., Elger, C.E., 2007. Epilepsy: accuracy of patient seizure counts. Archives of Neurology 64, 1595. doi:10.1001/archneur.64.11.1595.

[53] Hossain, M.R.I., Little, J.J., 2018. Exploiting temporal information for 3D human pose estimation, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 69–86. doi:10.1007/978-3-030-01249-6_5.

[54] Hou, J.C., McGonigal, A., Bartolomei, F., Thonnat, M., 2021. A Multi-Stream Approach for Seizure Classification with Knowledge Distillation. AVSS 2021 - 17th IEEE International Conference on Advanced Video and Signal-Based Surveillance doi:10.1109/AVSS52988.2021.9663770.

[55] Hou, J.C., McGonigal, A., Bartolomei, F., Thonnat, M., 2022. A Self-Supervised Pre-Training Framework for Vision-Based Seizure Classification , 1151–1155doi:10.1109/ICASSP43922.2022.9746325.

[56] Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C., 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. Technical Report. URL: http://vision.imar.ro/human3.6m.

[57] Iskakov, K., Burkov, E., Lempitsky, V., Malkov, Y., 2019. Learnable triangulation of human pose, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 7717–7726. URL: https://saic-violet.github.io/, doi:10.1109/ICCV.2019.00781.

[58] Jahangiri, E., Yuille, A.L., 2017. Generating Multiple Diverse Hypotheses for Human 3D Pose Consistent with 2D Joint Detections, in: Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017, pp. 805–814. doi:10.1109/ICCVW.2017.100.

[59] Jiang, L., Wang, N., Dang, Q., Liu, R., Lai, B., 2021. PP-MSVSR: Multi-Stage Video Super-Resolution. arXiv:2112.02828v1 URL: https://arxiv.org/abs/2112.02828v1http://arxiv.org/abs/2112.02828.

[60] Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y., 2015. Panoptic Studio: A Massively Multiview System for Social Motion Capture *. ICCV 2015 URL: http://www.cs.cmu.edu/.

[61] Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y., 2019. Panoptic Studio: A Massively Multiview System for Social Interaction Capture. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 190–204. doi:10.1109/TPAMI.2017.2782743.

[62] Kadkhodamohammadi, A., Padoy, N., 2021. A generalizable approach for multi-view 3D human pose regression. Machine Vision and Applications 32. doi:10.1007/s00138-020-01120-2.

[63] Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J., 2018. End-to-End Recovery of Human Shape and Pose, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 7122–7131. URL: https://akanazawa.github.io/hmr/., doi:10.1109/CVPR.2018.00744.

[64] Karacsony, T., Loesch-Biffar, A.M., Vollmar, C., Noachtar, S., Cunha, J.P.S., 2020. A Deep Learning Architecture for Epileptic Seizure Classification Based on Object and Action Recognition, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Institute of Electrical and Electronics Engineers (IEEE), Barcelona, Spain. pp. 4117–4121. doi:10.1109/icassp40776.2020.9054649.

[65] Karácsony, T., Loesch-Biffar, A.M., Vollmar, C., Noachtar, S., Cunha, J.P.S., 2021. Deepepil: Towards an Epileptologist-Friendly AI Enabled Seizure Classification Cloud System based on Deep Learning Analysis of 3D videos. BHI 2021 - 2021 IEEE EMBS International Conference on Biomedical and Health Informatics, Proceedings doi:10.1109/BHI50953.2021.9508555.

[66] Karácsony, T., Loesch-Biffar, A.M., Vollmar, C., Rémi, J., Noachtar, S., Cunha, J.P.S., 2022. Novel 3D video action recognition deep learning approach for near real time epileptic seizure classification. Scientific Reports 2022 12:1 12, 1–13. URL: https://www.nature.com/articles/s41598-022-23133-9, doi:10.1038/s41598-022-23133-9.

[67] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A., 2017. The Kinetics Human Action Video Dataset. arXiv:1705.06950 URL: http://deepmind.http://arxiv.org/abs/1705.06950.

[68] Kerling, F., Mueller, S., Pauli, E., Stefan, H., 2006. When do patients forget their seizures? An electroclinical study. Epilepsy & Behavior 9, 281–285. doi:10.1016/j.yebeh.2006.05.010.

[69] Kocabas, M., Athanasiou, N., Black, M.J., 2020. Vibe: Video inference for human body pose and shape estimation, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 5252–5262. URL: https://github.com/

mkocabas/VIBE, doi:10.1109/CVPR42600.2020.00530.

[70] Kocabas, M., Huang, C.H.P., Tesch, J., Müller, L., Hilliges, O., Black, M.J., 2021. SPEC: Seeing People in the Wild with an Estimated Camera. ICCV 2021 URL: https://spec.is.tue.mpg.de/.http://arxiv.org/abs/2110.00620.

[71] Kolotouros, N., Pavlakos, G., Black, M., Daniilidis, K., 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2252–2261. URL: https://seas.upenn.edu/, doi:10.1109/ICCV.2019.00234.

[72] Kondratyuk, D., Yuan, L., Li, Y., Zhang, L., Tan, M., Brown, M., Gong, B., 2021. Movinets: Mobile video networks for efficient video recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 16015–16025. URL: https://github.com/tensorflow/models/, doi:10.1109/CVPR46437.2021.01576.

[73] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T., 2011. HMDB: A large video database for human motion recognition, in: 2011 International Conference on Computer Vision, IEEE. doi:10.1109/iccv.2011.6126543.

[74] Kumar, M., Weissenborn, D., Kalchbrenner, N., 2021. Colorization Transformer. International Conference on Learning Representations URL: http://arxiv.org/abs/2102.04432.

[75] Kundu, J.N., Seth, S., Jampani, V., Rakesh, M., Venkatesh Babu, R., Chakraborty, A., 2020a. Self-supervised 3D human pose estimation via part guided novel image synthesis, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 6151–6161. URL: http://val.cds.iisc.ac.in/pgp-human/, doi:10.1109/CVPR42600.2020.00619.

[76] Kundu, J.N., Seth, S., Rahul, M.V., Rakesh, M., Babu, R.V., Chakraborty, A., 2020b. Kinematic-Structure-Preserved Representation for Unsupervised 3D Human Pose Estimation. Proceedings of the AAAI Conference on Artificial Intelligence 34, 11312–11319. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6792, doi:10.1609/AAAI.V34I07.6792.

[77] Kurada, A.V., Srinivasan, T., Hammond, S., Ulate-Campos, A., Bidwell, J., 2019. Seizure detection devices for use in antiseizure medication clinical trials: A systematic review. Seizure 66, 61–69. doi:10.1016/j.seizure.2019.02.007.

[78] Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V., 2017. Unite the people: Closing the loop between 3D and 2D human representations, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 4704–4713. doi:10.1109/CVPR.2017.500.

[79] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W., 2017. Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 105–114. doi:10.1109/CVPR.2017.19.

[80] Lee, K., Lee, I., Lee, S., 2018. Propagating LSTM: 3D pose estimation based on joint interdependency, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 123–141. doi:10.1007/978-3-030-01234-2_8.

[81] Li, C., Lee, G.H., 2019. Generating multiple hypotheses for 3D human pose estimation with mixture density network, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 9879–9887. URL: https://github.com/chaneyddt/Generating-, doi:10.1109/CVPR.2019.01012.

[82] Li, J., Liang, X., Wei, Y., Xu, T., Feng, J., Yan, S., 2017. Perceptual generative adversarial networks for small object detection, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 1951–1959. doi:10.1109/CVPR.2017.211.

[83] Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q., 2022a. Symbiotic Graph Neural Networks for 3D Skeleton-Based Human Action Recognition and Motion Prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 3316–3333. doi:10.1109/TPAMI.2021.3053765.

[84] Li, W., Liu, H., Ding, R., Liu, M., Wang, P., Yang, W., 2022b. Exploiting Temporal Contexts with Strided Transformer for 3D Human Pose Estimation. IEEE Transactions on Multimedia URL: https://github.com/Vegetebird/StridedTransformer-Pose3D., doi:10.1109/TMM.2022.3141231.

[85] Li, Y., Jin, P., Yang, F., Liu, C., Yang, M.H., Milanfar, P., 2021a. COMISR: Compression-Informed Video Super-Resolution. ICCV IEEE International Conference on Computer Vision .

[86] Li, Y., Sixou, B., Peyrin, F., 2021b. A Review of the Deep Learning Methods for Medical Images Super Resolution Problems. doi:10.1016/j.irbm.2020.08.004.

[87] Li, Z., Wang, X., Wang, F., Jiang, P., 2019. On boosting single-frame 3D human pose estimation via monocular videos, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2192–2201. doi:10.1109/ICCV.2019.00228.

[88] Liang, J., Cao, J., Fan, Y., Zhang, K., Ranjan, R., Li, Y., Timofte, R., Van Gool, L., 2022. VRT: A Video Restoration Transformer. arXiv:2201.12288v1 URL: https://github.com/JingyunLiang/VRThttp://arxiv.org/abs/2201.12288.

[89] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R., 2021. SwinIR: Image Restoration Using Swin Transformer, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1833–1844. URL: https://github.com/JingyunLiang/SwinIR, doi:10.1109/ICCVW54120.2021.00210.

[90] Liang, S., Sun, X., Wei, Y., 2018. Compositional Human Pose Regression. Computer Vision and Image Understanding 176-177, 1–8. doi:10.1016/j.cviu.2018.10.006.

[91] Lin, K., Wang, L., Liu, Z., 2021a. End-to-End Human Pose and Mesh Reconstruction with Transformers, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1954–1963. doi:10.1109/CVPR46437.2021.00199.

[92] Lin, K., Wang, L., Liu, Z., 2021b. Mesh Graphormer. ICCV 2021 URL: https://github.http://arxiv.org/abs/2104.00272.

[93] Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C., 2020a. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence 42, 2684–2701. doi:10.1109/TPAMI.2019.2916873.

[94] Liu, K., Ding, R., Zou, Z., Wang, L., Tang, W., 2020b. A Comprehensive Study of Weight Sharing in Graph Networks for 3D Human Pose Estimation, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 318–334. doi:10.1007/978-3-030-58607-2_19.

[95] Liu, S., Huang, X., Fu, N., Li, C., Su, Z., Ostadabbas, S., 2022. Simultaneously-Collected Multimodal Lying Pose Dataset: Enabling In-Bed Human Pose Monitoring. IEEE Transactions on Pattern Analysis and Machine Intelligence doi:10.1109/TPAMI.2022.3155712.

[96] Liu, S., Ostadabbas, S., 2019. Seeing under the cover: A physics guided learning approach for in-bed pose estimation, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer. pp. 236–245. URL: https://link.springer.com/chapter/10.1007/978-3-030-32239-7_27, doi:10.1007/978-3-030-32239-7_27.

[97] Liu, T., Sun, J.J., Zhao, L., Zhao, J., Yuan, L., Wang, Y., Chen, L.C., Schroff, F., Adam, H., 2021. View-Invariant, Occlusion-Robust Probabilistic Embedding for Human Pose. International Journal of Computer Vision URL: https://github.com/google-, doi:10.1007/s11263-021-01529-w.

[98] Liu, W., Mei, T., 2022. Recent Advances of Monocular 2D and 3D Human Pose Estimation: A Deep Learning Perspective. ACM Computing Surveys URL: https://doi.org/10.1145/3524497, doi:10.1145/3524497.

[99] Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W., 2020c. Disentangling and unifying graph convolutions for skeleton-based action recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 140–149. doi:10.1109/CVPR42600.2020.00022.

[100] Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M., 2019. AMASS: Archive of motion capture as surface shapes, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 5441–5450. doi:10.1109/ICCV.2019.00554.

[101] Maia, P., Hartl, E., Vollmar, C., Noachtar, S., Cunha, J.P.S., 2019. Epileptic seizure classification using the NeuroMov database, in: 2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG), IEEE. doi:10.1109/enbeng.2019.8692465.

[102] von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G., 2018. Recovering accurate 3D human pose in the wild using IMUs and a moving camera, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 614–631. URL: http://virtualhumans.mpi-inf.mpg.de/3DPW., doi:10.1007/978-3-030-01249-6_37.

[103] Martinez, J., Hossain, R., Romero, J., Little, J.J., 2017. A Simple Yet Effective Baseline for 3d Human Pose Estimation, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2659–2668. URL: https://github.com/una-dinosauria/, doi:10.1109/ICCV.2017.288.

[104] Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C., 2018. Monocular 3D human pose estimation in the wild using improved CNN supervision, in: Proceedings - 2017 International Conference on 3D Vision, 3DV 2017, pp. 506–516. doi:10.1109/3DV.2017.00064.

[105] Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Elgharib, M., Fua, P., Seidel, H.P., Rhodin, H., Pons-Moll, G., Theobalt, C., 2020. XNect: Real-time Multi-Person 3D Motion Capture with a Single RGB Camera. ACM Transactions on Graphics 39. URL: https://doi.org/10.1145/3386569.3392410., doi:10.1145/3386569.3392410.

[106] Moreno-Noguer, F., 2017. 3D human pose estimation from a single image via distance matrix regression, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 1561–1570. doi:10.1109/CVPR.2017.170.

[107] Neumann, L., Vedaldi, A., 2018. Tiny People Pose. ACCV .

[108] Nie, B.X., Wei, P., Zhu, S.C., 2017. Monocular 3D Human Pose Estimation by Predicting Depth on Joints, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3467–3475. doi:10.1109/ICCV.2017.373.

[109] Nie, Q., Liu, Z., Liu, Y., 2020. Unsupervised 3D Human Pose Representation with Viewpoint and Pose Disentanglement, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 102–118. URL: https://github.com/NIEQiang001/unsupervised-human-pose.git., doi:10.1007/978-3-030-58529-7_7.

[110] Noachtar, S., Borggraefe, I., 2009. Epilepsy surgery: A critical review. Epilepsy & Behavior 15, 66–72. doi:10.1016/j.yebeh.2009.02.028.

[111] Noachtar, S., Peters, A.S., 2009. Semiology of epileptic seizures: A critical review. Epilepsy & Behavior 15, 2–9. doi:10.1016/j.yebeh.2009.02.029.

[112] Noh, J., Bae, W., Lee, W., Seo, J., Kim, G., 2019. Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 9724–9733. URL: http://vision.snu.ac.kr/projects/better-to-follow, doi:10.1109/ICCV.2019.00982.

[113] Oh, S.J., Murphy, K., Pan, J., Roth, J., Schroff, F., Gallagher, A., 2019. Modeling uncertainty with hedged instance embedding, in: 7th International Conference on Learning Representations, ICLR 2019. URL: https://github.com/google/n-digit-mnist.

[114] Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., Schiele, B., 2018. Neural body fitting: Unifying deep learning and model based human pose and shape estimation, in: Proceedings - 2018 International Conference on 3D Vision, 3DV 2018, pp. 484–494. URL: http://github.com/mohomran/, doi:10.1109/3DV.2018.00062.

[115] Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation Learning with Contrastive Predictive Coding. Advances in neural information processing systems URL: http://arxiv.org/abs/1807.03748.

[116] Pandey, K., Mukherjee, A., Rai, P., Kumar, A., 2022. DiffuseVAE: Efficient, Controllable and High-Fidelity Generation from Low-Dimensional Latents URL: https://github.com/kpandey008/DiffuseVAE.http://arxiv.org/abs/2201.00308.

[117] Pareek, P., Thakkar, A., 2021. A survey on video-based Human Action Recognition: recent updates, datasets, challenges, and applications. Artificial Intelligence Review 54, 2259–2322. URL: https://link.springer.com/article/10.1007/s10462-020-09904-8, doi:10.1007/S10462-020-09904-8/TABLES/13.

[118] Patel, P., Huang, C.H.P., Tesch, J., Hoffmann, D.T., Tripathi, S., Black, M.J., 2021. AGORA: Avatars in Geography Optimized for Regression Analysis, pp. 13463–13473. URL: https://agora.is.tue.mpg.de/., doi:10.1109/cvpr46437.2021.01326.

[119] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J., 2019. Expressive body capture: 3D hands, face, and body from a single image, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 10967–10977. URL: https://smpl-x.is.tue.mpg.de., doi:10.1109/CVPR.2019.01123.

[120] Pavlakos, G., Zhou, X., Daniilidis, K., 2018a. Ordinal Depth Supervision for 3D Human Pose Estimation, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 7307–7316. doi:10.1109/CVPR.2018.00763.

[121] Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K., 2017. Coarse-to-fine volumetric prediction for single-image 3D human pose, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 1263–1272. doi:10.1109/CVPR.2017.139.

[122] Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K., 2018b. Learning to Estimate 3D Human Pose and Shape from a Single Color Image, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 459–468. doi:10.1109/CVPR.2018.00055.

[123] Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M., 2019. 3D human pose estimation in video with temporal convolutions and semi-supervised training, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 7745–7754. URL: https://github.com/, doi:10.1109/CVPR.2019.00794.

[124] Peng Guan, Weiss, A., Balan, A.O., Black, M.J., 2010. Estimating human shape and pose from a single image, Institute of Electrical and Electronics Engineers (IEEE). pp. 1381–1388. doi:10.1109/iccv.2009.5459300.

[125] Pérez-García, F., Scott, C., Sparks, R., Diehl, B., Ourselin, S., 2021. Transfer Learning of Deep Spatiotemporal Networks to Model Arbitrarily Long Videos of Seizures. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12905 LNCS, 334–344. URL: https://link.springer.com/chapter/10.1007/978-3-030-87240-3_32, doi:10.1007/978-3-030-87240-3_32.

[126] Pothula, P.K., Marisetty, S., Rao, M., 2022. A Real-Time Seizure Classification System Using Computer Vision Techniques. 2022 IEEE International Systems Conference (SysCon) , 1–6URL: https://ieeexplore.ieee.org/document/9773923/, doi:10.1109/SYSCON53536.2022.9773923.

[127] Qammaz, A., Argyros, A., 2020. Occlusion-tolerant and personalized 3D human pose estimation in RGB images, in: Proceedings - International Conference on Pattern Recognition, Institute of Electrical and Electronics Engineers Inc.. pp. 6904–6911. doi:10.1109/ICPR48806.2021.9411956.

[128] Reddy, N.D., Guigues, L., Pishchulin, L., Eledath, J., Narasimhan, S.G., 2021. TesseTrack: End-to-End Learnable Multi-Person Articulated 3D Pose Tracking, in: CVPR2021, pp. 15185–15195. URL: http://www.cs.cmu.edu/, doi:10.1109/cvpr46437.2021.01494.

[129] Reddy, N.D., Vo, M., Narasimhan, S.G., 2018. CarFusion: Combining Point Tracking and Part Detection for Dynamic 3D Reconstruction of Vehicles, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1906–1915. doi:10.1109/CVPR.2018.00204.

[130] Ren, W., Ma, O., Ji, H., Liu, X., 2020. Human Posture Recognition Using a Hybrid of Fuzzy Logic and Machine Learning Approaches. IEEE Access 8, 135628–135639. doi:10.1109/ACCESS.2020.3011697.

[131] Rong, Y., Shiratori, T., Joo, H., 2020. FrankMocap: Fast Monocular 3D Hand and Body Motion Capture by Regression and Integration. arXiv URL: http://arxiv.org/abs/2008.08324.

[132] Rosenow, F., 2001. Presurgical Evaluation of Epilepsy. Brain 124, 1683–1700. doi:10.1093/brain/124.9.1683.

[133] Rozumnyi, D., Oswald, M.R., Ferrari, V., Matas, J., Pollefeys, M., 2021. DeFMO: Deblurring and Shape Recovery of Fast Moving Objects. IEEE Computer Society Conference on Computer Vision and Pattern Recognition .

[134] Sárándi, I., Linder, T., Arras, K.O., Leibe, B., 2018. How Robust is 3D Human Pose Estimation to Occlusion? URL: http://arxiv.org/abs/1808.09316.

[135] Schroff, F., Kalenichenko, D., Philbin, J., 2015. FaceNet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 815–823. doi:10.1109/CVPR.2015.7298682.

[136] Shaham, T.R., Dekel, T., Michaeli, T., 2019. SinGAN: Learning a generative model from a single natural image, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 4569–4579. doi:10.1109/ICCV.2019.00467.

[137] Shahroudy, A., Liu, J., Ng, T.T., Wang, G., 2016. NTU RGB+D: A large scale dataset for 3D human activity analysis, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1010–1019. doi:10.1109/CVPR.2016.115.

[138] Sharma, S., Varigonda, P.T., Bindal, P., Sharma, A., Jain, A., 2019. Monocular 3D human pose estimation by generation and ordinal ranking, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2325–2334. URL: https://github.com/, doi:10.1109/ICCV.2019.00241.

[139] Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z., 2016. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1874–1883. doi:10.1109/CVPR.2016.207.

[140] Sigal, L., Balan, A.O., Black, M.J., 2010. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International Journal of Computer Vision 87, 4–27. URL: http://www.cs.brown.edu/people/ls/ehum2/., doi:10.1007/s11263-009-0273-6.

[141] Smaira, L., Carreira, J., Noland, E., Clancy, E., Wu, A., Zisserman, A., 2020. A Short Note on the Kinetics-700-2020 Human Action Dataset. arXiv:1907.06987v1 URL: https://www.babbel.com/en/magazine/the-10-most-http://arxiv.org/abs/2010.10864.

[142] Song, Y.F., Zhang, Z., Shan, C., Wang, L., 2022. Constructing Stronger and Faster Baselines for Skeleton-based Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence URL: https://github.com/yfsong0709/EfficientGCNv1., doi:10.1109/TPAMI.2022.3157033.

[143] Soomro, K., Zamir, A.R., Shah, M., 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. arXiv:1212.0402 URL: https://arxiv.org/abs/1212.0402v1http://arxiv.org/abs/1212.0402, doi:10.48550/arxiv.1212.0402.

[144] Srivastav, V., Issenhuth, T., Kadkhodamohammadi, A., de Mathelin, M., Gangi, A., Padoy, N., 2018. MVOR: A Multi-view RGB-D Operating Room Dataset for 2D and 3D Human Pose Estimation. MICCAI-LABELS URL: http://camma.u-strasbg.fr/datasetshttp://arxiv.org/abs/1808.08180.

[145] Su, J.W., Chu, H.K., Huang, J.B., 2020. Instance-aware image colorization, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 7965–7974. URL: https://cgv.cs.nthu.edu.tw/projects/instaColorization, doi:10.1109/CVPR42600.2020.00799.

[146] Sun, W., Chen, Z., 2019. Learned Image Downscaling for Upscaling using Content Adaptive Resampler. IEEE Transactions on Image Processing 29, 4027–4040. URL: http://arxiv.org/abs/1907.12904http://dx.doi.org/10.1109/TIP.2020.2970248, doi:10.1109/TIP.2020.2970248.

[147] Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., Liu, J., 2022. Human Action Recognition From Various Data Modalities: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence , 1–20URL: https://ieeexplore.ieee.org/document/9795869/, doi:10.1109/TPAMI.2022.3183112.

[148] Tan, J.K.V., Budvytis, I., Cipolla, R., 2017. Indirect deep structured learning for 3D human body shape and pose prediction, in: British Machine Vision Conference 2017, BMVC 2017. doi:10.5244/c.31.15.

[149] Tan, W., Yan, B., Bare, B., 2018. Feature Super-Resolution: Make Machine See More Clearly Feature Super-Resolution Image Super-Resolution High Discriminative Feature low-resolution image high-resolution image Viewing Recognition (a) Feature Super-Resolution: a novel super-resolution appro. Computer Vision and Pattern Recognition (CVPR) , 4321–4329URL: http://openaccess.thecvf.com/content_cvpr_2018/papers/Tan_Feature_Super-Resolution_Make_CVPR_2018_paper.pdf.

[150] Tekin, B., Katircioglu, I., Salzmann, M., Lepetit, V., Fua, P., 2016a. Structured prediction of 3D human pose with deep neural networks, in: British Machine Vision Conference 2016, BMVC 2016, pp. 1–130. doi:10.5244/C.30.130.

[151] Tekin, B., Marquez-Neila, P., Salzmann, M., Fua, P., 2017. Learning to Fuse 2D and 3D Image Cues for Monocular Body Pose Estimation, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3961–3970. doi:10.1109/ICCV.2017.425.

[152] Tekin, B., Rozantsev, A., Lepetit, V., Fua, P., 2016b. Direct prediction of 3D body poses from motion compensated sequences, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 991–1000. doi:10.1109/CVPR.2016.113.

[153] Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J., 2017. Total capture: 3D human pose estimation fusing video and inertial sensors, in: British Machine Vision Conference 2017, BMVC 2017. URL: http://cvssp.org/data/totalcapture/., doi:10.5244/c.31.14.

[154] Tung, H.Y.F., Tung, H.W., Yumer, E., Fragkiadaki, K., 2017. Self-supervised learning of motion capture, in: Advances in Neural Information Processing Systems, pp. 5237–5247.

[155] Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C., 2017. Learning from synthetic humans, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Institute of Electrical and Electronics Engineers Inc.. pp. 4627–4635. doi:10.1109/CVPR.2017.492.

[156] Wang, C., Zhang, F., Zhu, X., Ge, S.S., 2021a. Low-resolution Human Pose Estimation URL: http://arxiv.org/abs/2109.09090.

[157] Wang, J., Huang, S., Wang, X., Tao, D., 2019. Not all parts are created equal: 3D pose estimation by modeling bi-directional dependencies of body parts, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 7770–7779. doi:10.1109/ICCV.2019.00786.

[158] Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y., 2014. Learning fine-grained image similarity with deep ranking, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1386–1393. URL: https://sites.google.com/site/imagesimilaritydata/., doi:10.1109/CVPR.2014.180.

[159] Wang, J., Yan, S., Xiong, Y., Lin, D., 2020a. Motion Guided 3D Pose Estimation from Videos, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 764–780. URL: https://www.youtube.com/watch?v=VHhsXG6OXnI&t=87s., doi:10.1007/978-3-030-58601-0_45.

[160] Wang, M., Chen, X., Liu, W., Qian, C., Lin, L., Ma, L., 2018a. Drpose3D: Depth ranking in 3D human pose estimation. IJCAI International Joint Conference on Artificial Intelligence 2018-July, 978–984. doi:10.24963/IJCAI.2018/136.

[161] Wang, P., Li, W., Ogunbona, P., Wan, J., Escalera, S., 2018b. RGB-D-based human motion recognition with deep learning: A survey. Computer Vision and Image Understanding 171, 118–139. doi:10.1016/j.cviu.2018.04.007.

[162] Wang, X., Gao, L., Wang, P., Sun, X., Liu, X., 2018c. Two-Stream 3-D convNet Fusion for Action Recognition in Videos With Arbitrary Size and Length. IEEE Transactions on Multimedia 20, 634–644. URL: https://dl.acm.org/doi/abs/10.1109/TMM.2017.2749159, doi:10.1109/TMM.2017.2749159.

[163] Wang, Z., Chang, S., Yang, Y., Liu, D., Huang, T.S., 2016. Studying very low resolution recognition using deep networks, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 4792–4800. doi:10.1109/CVPR.2016.518.

[164] Wang, Z., Chen, J., Hoi, S.C., 2021b. Deep Learning for Image Super-Resolution: A Survey. doi:10.1109/TPAMI.2020.2982166.

[165] Wang, Z., Shin, D., Fowlkes, C.C., 2020b. Predicting Camera Viewpoint Improves Cross-Dataset Generalization for 3D Human Pose Estimation, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Science and Business Media Deutschland GmbH. pp. 523–540. doi:10.1007/978-3-030-66096-3_36.

[166] Wohlhart, P., Lepetit, V., 2015. Learning descriptors for object recognition and 3D pose estimation, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3109–3118. doi:10.1109/CVPR.2015.7298930.

[167] Xiang, D., Joo, H., Sheikh, Y., 2019. Monocular total capture: Posing face, body, and hands in the wild, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 10957–10966. doi:10.1109/CVPR.2019.01122.

[168] Xie, K., Wang, T., Iqbal, U., Guo, Y., Fidler, S., Shkurti, F., 2021. Physics-based Human Motion Estimation and Synthesis from Videos, in: ICCV2021. URL: http://arxiv.org/abs/2109.09913.

[169] Xu, J., Yu, Z., Ni, B., Yang, J., Yang, X., Zhang, W., 2020a. Deep Kinematics Analysis for Monocular 3D Human Pose Estimation, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 896–905. doi:10.1109/CVPR42600.2020.00098.

[170] Xu, X., Chen, H., Moreno-Noguer, F., Jeni, L.A., De la Torre, F., 2020b. 3D Human Shape and Pose from a Single Low-Resolution Image with Self-Supervised Learning, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 284–300. doi:10.1007/978-3-030-58545-7_17.

[171] Xu, X., Chen, H., Moreno-Noguer, F., Jeni, L.A., De la Torre, F., 2021. 3D Human Pose, Shape and Texture from Low-Resolution Images and Videos. IEEE Transactions on Pattern Analysis and Machine Intelligence doi:10.1109/TPAMI.2021.3070002.

[172] Yan, S., Xiong, X., Arnab, A., Lu, Z., Zhang, M., Sun, C., Schmid, C., 2022. Multiview Transformers for Video Recognition. CVPR2022 URL: https://github.com/google-research/scenic.http://arxiv.org/abs/2201.04288.

[173] Yan, S., Xiong, Y., Lin, D., 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. 32nd AAAI Conference on Artificial Intelligence, AAAI 2018 , 7444–7452URL: https://arxiv.org/abs/1801.07455v2, doi:10.48550/arxiv.1801.07455.

[174] Yin, Y., Robinson, J.P., Fu, Y., 2020. Multimodal In-bed Pose and Shape Estimation under the Blankets. arXiv:2012.06735 URL: http://arxiv.org/abs/2012.06735.

[175] Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y., 2022. CoCa: Contrastive Captioners are Image-Text Foundation Models. arXiv:2205.01917 URL: http://arxiv.org/abs/2205.01917.

[176] Yu, T., Guo, K., Xu, F., Dong, Y., Su, Z., Zhao, J., Li, J., Dai, Q., Liu, Y., 2017. BodyFusion: Real-Time Capture of Human Motion and Surface Geometry Using a Single Depth Camera, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 910–919. doi:10.1109/ICCV.2017.104.

[177] Yu, Y., Fan, H., Zhang, L., 2023. Deficiency-aware masked transformer for video inpainting. arXiv preprint arXiv:2307.08629 .

[178] Zanfir, A., Marinoiu, E., Sminchisescu, C., 2018. Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes: The Importance of Multiple Scene Constraints, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2148–2157. doi:10.1109/CVPR.2018.00229.

[179] Zavala-Mondragon, L.A., Lamichhane, B., Zhang, L., Haan, G.d., 2020. CNN-SkelPose: a CNN-based skeleton estimation algorithm for clinical applications. Journal of Ambient Intelligence and Humanized Computing 11, 2369–2380. doi:10.1007/s12652-019-01259-5.

[180] Zeng, A., Sun, X., Huang, F., Liu, M., Xu, Q., Lin, S., 2020. SRNet: Improving Generalization in 3D Human Pose Estimation with a Split-and-Recombine Approach, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 507–523. doi:10.1007/978-3-030-58568-6_30.

[181] Zhang, S., Zhang, Y., Bogo, F., Pollefeys, M., Tang, S., Zürich, E., 2021. Learning Motion Priors for 4D Human Body Capture in 3D Scenes. Iccv URL: https://sanweiliti.

[182] Zhang, Z., Wang, C., Qin, W., Zeng, W., 2020. Fusing wearable IMUs with multi-view images for human pose estimation: A geometric approach, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2197–2206. doi:10.1109/CVPR42600.2020.00227.

[183] Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N., 2019. Semantic graph convolutional networks for 3D human pose regression, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3420–3430. doi:10.1109/CVPR.2019.00354.

[184] Zheng, C., Wu, W., Yang, T., Zhu, S., Chen, C., Liu, R., Shen, J., Kehtarnavaz, N., Shah, M., 2022. Deep learning-based human pose estimation: A survey. arXiv preprint arXiv:2012.13392 URL: https://arxiv.org/abs/2012.13392.

[185] Zhou, K., Han, X., Jiang, N., Jia, K., Lu, J., 2021. HEMlets PoSh: Learning Part-Centric Heatmap Triplets for 3D Human Pose and Shape Estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence doi:10.1109/TPAMI.2021.3051173.

[186] Zhou, S., Li, C., Chan, K.C., Loy, C.C., 2023. Propainter: Improving propagation and transformer for video inpainting, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10477–10486.

[187] Zhou, X., Sun, X., Zhang, W., Liang, S., Wei, Y., 2016. Deep kinematic pose regression, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 186–201. doi:10.1007/978-3-319-49409-8_17.

[188] Zhou, X., Zhu, M., Pavlakos, G., Leonardos, S., Derpanis, K.G., Daniilidis, K., 2019. Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. IEEE Transactions on Pattern Analysis; Machine Intelligence 41, 901–914. doi:10.1109/TPAMI.2018.2816031.