SPATIOTEMPORAL CHARACTERIZATION OF GAIT FROM MONOCULAR VIDEOS WITH TRANSFORMERS

R. James Cotton[1,2], Emoonah McClerklin[1], Anthony Cimorelli[1], Ankit Patel[3], Tasos Karakostas[1] *

ABSTRACT

Human pose estimation from monocular video is a rapidly advancing field that offers great promise to human movement science and rehabilitation. This potential is tempered by the smaller body of work ensuring the outputs are clinically meaningful and properly calibrated. Gait analysis, typically performed in a dedicated lab, produces precise measurements including kinematics and step timing. Using more than 9000 monocular video from an instrumented gait analysis lab, we evaluated the performance of existing algorithms for measuring kinematics. While they produced plausible results that resemble walking, the joint angles and step length were noisy and poorly calibrated. We trained a transformer to map 3D joint location sequences and the height of individuals onto interpretable biomechanical outputs including joint kinematics and phase within the gait cycle. This task-specific layer greatly reduced errors in the kinematics of the hip, knee and foot, and accurately detected the timing of foot down and up events. We show, for the first time, that accurate spatiotemporal gait parameters including walking speed, step length, cadence, double support time, and single support time can be computed on a cycle-by-cycle basis from these interpretable outputs. Our results indicate lifted 3D joint locations contain enough information for gait analysis, but their representation is not biomechanically accurate enough to use directly, suggesting room for improvement in existing algorithms[1].

1 INTRODUCTION

The remarkable progress in human pose estimation (HPE) from images and video offers great promise to human movement science and rehabilitation. State-of-the-art approaches enable high quality tracking of individuals in video and estimation of their joint locations – both in the 2D image plane and lifted to 3D coordinates (Zheng et al., 2020). However, the clinical utility of these algorithms are limited for several reasons (Seethapathi et al., 2019). Firstly, tools that produce clinically relevant measures of human movement are less common. HPE methods are typically trained to optimize the accuracy for estimating joint locations in Euclidean space. However, movements are rarely described this way. Rather, they are described by changes in joint angles following standard conventions (Wu et al., 2002; 2005). Furthermore, many activities are typically described higher levels that capture coordination over multiple joints (e.g., walking can be described by step length and frequency). Secondly, public datasets for HPE contain largely able-bodied individuals and how methods trained on these datasets perform when applied to patient populations has not been well studied. In general, AI fairness for people with disabilities has received relatively little attention (Trewin et al., 2019). In the context of HPE, methods may generalize poorly due to anatomical and movement pattern differences and the absence of assistive mobility devices and bracing in the training data, for example.

Gait impairments are common in rehabilitation (Verghese et al., 2006) and falls rank among the leading causes of death worldwide (World Health Organization, 2021). Gold standard clinical gait analysis is performed in a laboratory using optical motion capture and force plates to precisely measure joint angles, ground reaction forces, and the duration of different phases of gait as people walk (Richards et al., 2012). While these gait assessments provide precise measurements, the required

*1. Shirley Ryan AbilityLab, 2. Department of Physical Medicine and Rehabilitation, Northwestern University, 3. Department of Neuroscience, Baylor College of Medicine

[1]Code and model weights are available at (github link removed during anonymous review)

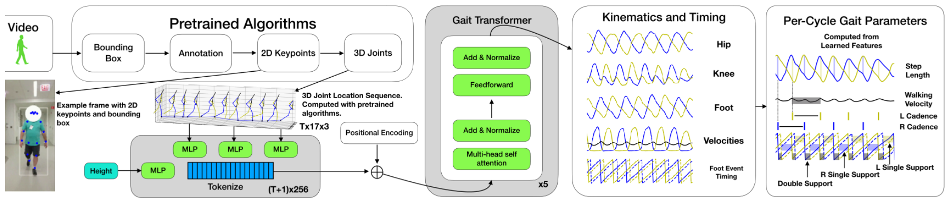

Figure 1: Overview of our gait analysis pipeline, with the components we train highlighted in gray. Video is first processed with published, pretrained algorithms to produce a sequence of 3D joint locations from video. A transformer is trained to produce interpretable kinematic parameters including hip and knee angles, foot position, and foot event timing. From these outputs, gait parameters can be computed from individual gait cycles.

equipment and expertise makes frequent, routine assessments impractical. A validated method to estimate a subset of the commonly measured gait parameters from monocular video would have significant clinical utility. For example, detection of gait impairment progression from video acquired at home may allow earlier intervention prior to falls.

Three-dimensional joint locations estimated from video are impressively accurate on several public datasets, with errors commonly below 50mm (Pavllo et al., 2019; Liu et al., 2020). To the best of our knowledge, the accuracy of kinematics (i.e. joint angles) computed from these joint locations during walking have not been evaluated, especially on clinical population, whom might benefit the most from these algorithms. We evaluated these algorithms on more than 9000 videos of 770 individuals walking who were seen for gait analysis. While computed 3D joint locations were plausible, the kinematics were noisy and poorly calibrated. We also found it challenging to accurately determine when the foot makes contact with the ground from the outputs, a prerequisite for defining the gait cycle and measuring most gait parameters.

These limitations motivated our approach (Fig. 1). We trained a transformer (Vaswani et al., 2017) on our dataset to map the 3D joint location sequences to kinematic trajectories and timing information, which also enabled accurate detection of foot down and up events. From these outputs, we measured several gait parameters on individual gait cycles and found a high correlation between our estimates and those from formal gait analysis. By producing meaningful trajectories from which we extract the gait parameters, our approach also is more explainable – a desirable feature in machine learning for medicine. Finally, we demonstrated that our approach generalizes to both Human 3.6 (Ionescu et al., 2014) and data collected in a clinic (i.e. outside the gait laboratory). The latter result is particularly relevant as it reflects our intended use case for a clinically adoptable gait analysis tool.

Our results reveal two critical points about the usability of HPE for gait analysis. (1) Despite being state of the art, 3D joint locations estimated via lifting do not produce an accurate enough representation to allow for directly computing relevant gait parameters and (2) that they *do* contain sufficient information to map them to the relevant interpretable biomechanics, which allows for accurately computing gait parameters. Ultimately, this last step is a patch to address the limitations of representations from existing algorithms and highlights opportunities for improvement.

Contributions Our main contributions are: ❶ We evaluate the accuracy of kinematics computed from 3D joint locations based on monocular video against ground-truth gait analysis laboratory data from patient populations and find they are noisy and poorly calibrated. ❷ We train a transformer to map these to less noisy and more accurate biomechanical trajectories and to the phase within the gait cycle. ❸ We show that these interpretable features allow accurate cycle-to-cycle estimates of common gait parameters, including cadence, walking velocity, step length, single support time, and double support time for each gait cycle. ❹ We show that our approach generalizes outside the training data, including the Human 3.6 dataset and videos acquired in an outpatient clinic.

2 Related Work

Human pose estimation We refer readers to Zheng et al. (2020); Liu et al. (2021) for an overview of the taxonomy of HPE and for a review of recent approaches and restrict ourselves to a brief discussion of the methods we use. Martinez et al. (2017) demonstrated lifting 2D keypoints to 3D joint locations achieved remarkable accuracy, which has subsequently been improved with techniques that process temporal sequences of 2D keypoints (Pavllo et al., 2019). In this work, we use Liu et al. (2020), which utilizes graph attention through time and over joints to further improve the accuracy of lifting. Lifting generalizes well as it it leverages the remarkable progress in 2D keypoint detection. Amongst these approaches, top-down methods that localize the joints of a person already identified by a bounding box achieve the greatest accuracy (e.g. Sun et al. (2019); Zhang et al. (2020a)). These are dependent on tracking algorithms that can identify a person throughout a video (e.g. Wojke & Bewley (2018); Zhang et al. (2020b)), which is also necessary in our project as there are commonly multiple people apparent in videos. These have also been advancing, with recent methods jointly trained to perform the detection of people and identification across frames (Zhang et al., 2020b; Sun et al., 2021).

One limitation of representing a body configuration with 3D joint locations is that inverse kinematics are required to recover the corresponding joint angles required. Methods that use parametric models of human bodies, such as HMR (Kanazawa et al., 2018), do not have this limitation as their outputs are joint angles. However, the accuracy of these methods does not reach 3D lifting and, often, inferred joint locations do not align with the images. We include experiments with VIBE (Kocabas et al., 2020), which maps an image sequence directly to a sequence of pose parameters and produces highly competitive performance on mesh recovery. Closely related to our focus on biomechanical accuracy of movements are methods that utilize physics simulations to produce more plausible movements (Yuan et al., 2021; Shimada et al., 2021; 2020; Shi et al., 2020). We did not begin with these methods as their reported joint accuracy is worse than lifting methods and most do not have available implementations, but we believe this to be a promising direction. These methods also detect foot-ground contacts, although the temporal accuracy of this was not reported – likely because ground truth timing was not collected in public datasets.

Gait Analysis Machine learning has been used to determine the timing of gait events from a number of sources including wearable sensors (Khera & Kumar, 2020). Neural networks have also been trained to detect the event times from motion capture data acquired at 120Hz, showing detection accuracy of 10ms for foot down and 13ms for foot up (Kidziński et al., 2019). Kanko et al. (2021a;b) have shown that multiple synchronized cameras enable accurate characterization of gait. Mehdizadeh et al. (2021) showed that 2D keypoints detected from monocular video allowed accurate measurement of cadence (steps per minute), although they did not quantify the temporal error for detecting the events and their system was less accurate for estimating step length. Kidziński et al. (2020) also used data from a gait laboratory and trained a neural network to predict walking speed and cadence from 2D keypoints sequences, although compared to our approach they produce an average parameter for a trajectory.

3 Gait Lab Dataset

Our dataset includes instrumented gait analysis from 770 subjects during 1073 sessions. This study was approved by the (removed during review) IRB. Subjects ages ranged from 2 to more than 80 with the median age being 11 years old and 90th percentile being 22 years old. Diagnoses documented during encounters ranged widely with cerebral palsy and spina bifida being common, and also included stroke, traumatic brain injury, spinal cord injury, amputation, and abnormality of gait. Sessions commonly involved gait analysis under multiple conditions (e.g. with or without a walker or brace, total number of unique conditions in the dataset are 2009) with several trials per condition. Each gait trial includes video acquired in the frontal plane (i.e. as the subject either walks towards or away from the camera) with synchronized motion capture data and force plate data acquired at 120Hz. Videos had previously been compressed to a resolution of 480×720 at 30fps. 615 subjects were randomized into the training set and 155 into the testing set.

Kinematic Trajectories, Gait Phases and Gait Parameters Our dataset was acquired during clinical practice and, as such, had previously been processed with a clinical workflow (Richards et al., 2012; Kadaba et al., 1990). This includes inverse kinematics solutions for individually-calibrated, anatomically-accurate biomechanical models to determine joint locations and angles from the surface marker locations. In this work, we only use kinematic trajectories in the sagit-

Parameter	Description
Cadence	Step frequency (steps / minutes), with two steps occurring per cycle
Step length	The forward distance between the feet when both are on the ground
Walking velocity	Forward movement of the pelvis over one gait cycle, divided by the duration
Double Stance Time	The duration within a gait cycle when both feet are on the ground
Single Support Time	The duration when only either the left or right foot is on the ground

Table 1: Description of spatiotemporal gait parameters estimated with our algorithm.

tal plane, including the flexion angles of the hip and knee, the forward position of the foot, and the forward velocity of the pelvis and feet (example traces are shown in Fig. 2). The time each foot goes up and down is detected by force plates and valid trials are required to have at least two down events for each foot. *We use the term kinematic trajectories and gait phases for these interpretable time-varying signals, which are the output of our model.*

A number of metrics are extracted from the kinematic trajectories and event times on each trial over a single gait cycle with respect to both the left and right foot. The ones we compute with our approach and analyze in this paper are described in Table 1. *We use the term gait parameters for these statistics, which are computed from the model outputs.*

4 Gait Analysis Pipeline

Video Processing Overview We restricted ourselves to pretrained algorithms to produce 3D joint locations as we worried training or fine tuning these steps on a uniform laboratory background of the videos from the frontal plane would limit the generalization outside of the laboratory or to new perspectives. Fig. 1 shows an overview of our pipelinee. (1) We first ran a tracking algorithm (Zhang et al., 2020b; Wojke & Bewley, 2018) to infer the bounding box tracks for all people in the scene followed by (2) manually annotating the bounding box for the person undergoing gait analysis. (3) Then we computed 2D keypoints (Zhang et al., 2020a; Sun et al., 2019), (4) and lifted them to 3D (Liu et al., 2020). We used DataJoint (Yatsenko et al., 2015) to manage the data and computational pipeline (Supplementary Figure 6). Please see appendix for more details.

Kinematics from 3D Joint Positions We computed kinematics from the lifted 3D joint locations. We focused on the accuracy of hip and knee flexion defined in the sagittal plane and the forward foot position because of their clinical relevance in many conditions. It is straightforward to obtain these from the 3D skeleton after rotating the skeleton around the vertical axis to align the vector from the left to right hip with the Y axis. Hip flexion angle is computed from the dot product between the vector from the spine joint to the mid-hip joint and the vector from the hip to the knee, both in the X-Z plane. The knee angle is computed from the dot product between the hip to knee vector and the knee to foot vector. We also extracted the position of each foot in the X (forward) axis, with the pelvis defined as the origin

Transformer for Interpretable Gait Features We trained a transformer (Vaswani et al., 2017) to map a sequence of lifted 3D joint locations and the height of the subject to a set of interpretable features for each frame. The kinematic outputs are the hip and knee joint angles and forward foot position for each side, described above, and additionally the forward velocity of the pelvis and each foot (for a total of 9 elements). The transformer also outputs timing information with respect to four gait events (left foot down, right foot down, left foot up, right foot up); we describe this representation below.

The 3D joint locations are tokenized as in Llopart (2020) by concatenating the joint locations for each frame, passing them through an MLP to match the embedding dimension, and using sinusoidal embedding for positional encoding. To include the subject's height, we provide a token using an additional MLP and a learned positional embedding. All of the tokens are concatenated and passed to the transformer.

Gait Phase Quadrature Encoding We represent the timing of the four periodic gait events by quadrature encoding the phase at each time point for each event, rather than directly predicting a sparse set of events or a binary output. We found the timing accuracy was similar with a binary output, this representation matches how gait analysis is normally described and facilitates aligning multiple cycles.

We computed the phase for all frames from two foot events in the dataset as $\phi_i(t) = 2\pi \frac{t-t_{i,0}}{t_{i,1}-t_{i,0}}$, where $t_{i,k}$ is the time of the k^{th} occurrence of gait event i. Some trials only had a single foot up event, in which case we replaced the denominator by the period between the same side down events. The phase was quadrature encoded as $\boldsymbol{q}_i(t) = [\cos\phi_i(t), \sin\phi_i(t)]$, which allows reconstructing the phase from the model outputs as:

$$\hat{\phi}_i(t) = \arctan\left(\boldsymbol{q}_{i,1}(t), \boldsymbol{q}_{i,0}(t)\right). \tag{1}$$

Sliding Window Inference We found the accuracy worsened when performing inference on longer sequences. We perform inference on longer sequences by applying the transformer to a sliding window of 3 seconds (90 frames) with stride of 1 frame and preserving the middle output frame for each window. Because the transformer produces a valid output corresponding to each input sample, we use this to pad the beginning and end of the output.

Data Augmentation All of our data is acquired from a frontal view (i.e. the person walking toward or away from the camera). To improve the generalization of the gait transformer to novel views, we augment the keypoints by applying a random rotation to the entire 3D keypoint sequence, with 50% probability. We do not perform any corresponding transformation to the outputs because the output format is viewpoint invariant.

Architecture Details and Training We refer readers to Vaswani et al. (2017) for most transformer details. Our encoder had 5 transformer layers with 6 attention heads in each layer, each with a dropout (Srivastava et al., 2014) probability of 0.1, and a projection dimension of 256. It was trained using an AdamW optimizer (Loshchilov & Hutter, 2019) for 250 epochs with a learning rate of $5e-4$ and weight decay of $1e-5$. Feed forward networks were a 2 layer MLP with 512 units in the first layer and using a GeLU (Hendrycks & Gimpel, 2016) nonlinearity followed by dropout layers with 0.1 probability. Layer normalization and layer scaling were both used (Ba et al., 2016; Touvron et al., 2021). Batches were grouped into buckets by length with batch sizes ranging from 128 for sequences of length 30 to 32 for length 300. The architecture and hyperparameters were selected when developing a precursor that only output the timing parameters using a small fraction of data and were not systematically explored. It had 9 million parameters and was implemented in TensorFlow 2 (Abadi et al., 2015) and trained in 1-2 hours on a 32GB A100 GPU. Sliding window inference on the testing data took approximately 4-5 hours.

The target gait phases extrapolate for times outside the two events and can become less accurate, so the loss used a weight of 1 between the two event times and a linear decay to zero by one second outside this range. The weight for the kinematic parameters was set to 0.1, and sequences were cropped to the time range when all the markers were tracked.

5 Experimental Results

Qualitative behavior of pretrained pipeline components on gait analysis subjects The most problematic step when processing the videos was the bounding box computation, and we noted several common problems. (1) When subjects came too close to the camera or briefly left the frame, they were not reliably reidentified. In these cases, our manual annotation tool allowed us to associate those tracks. (2) Many subjects required assistance with ambulation so there was frequently a therapist nearby, possibly making physical contact, which could result the bounding box track being fragmented by jumping from subject to therapist. In some cases identifiers would switch, in which case we would process with a different algorithm. (3) With FairMOT in particular, we noted the presence of a rolling walker significantly increased the chance that a person was not detected. This has significant implications for both AI fairness for people with disabilities and safety, when considering robotic applications that might fail to detect these individuals (Trewin et al., 2019).

We found that 2D keypoint detection with pretrained networks based on able-bodied populations generally performed well on rehabilitation patients, provided the bounding boxes were able to track

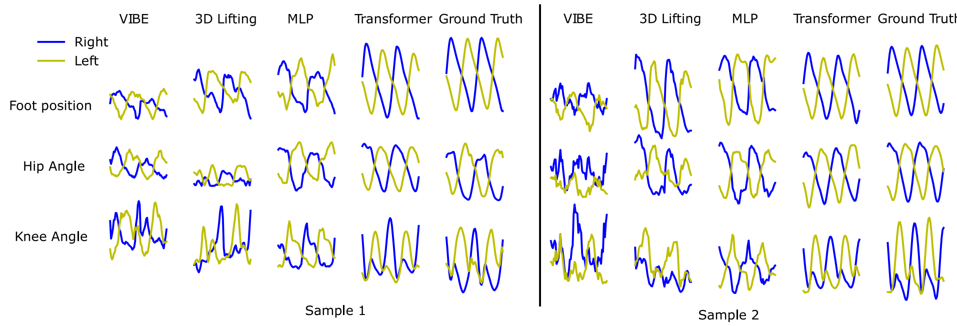

Figure 2: Two sample trajectories (left and right). Each row shows a different joint and each column is a different approach.

the individual. This included situations where a therapist was nearby or even behind the subject of interest, with the algorithms tracking the person in the foreground centered in the bounding box. We did note the presence of ankle-foot orthoses seems to worsen ankle localization and canes or crutches worsened upper extremity locations. Some frames would also briefly confuse the left and right sides, which would adversely impact the 3D joint locations and our algorithm's performance. The 3D joint locations produced plausible poses that visually resembled walking (e.g. see skeleton in Fig. 1) provided the 2D keypoints, although our quantitative results below highlight the limitations.

Failure modes of VIBE (Kocabas et al., 2020) were more dramatic, with flipping of the direction of the person. This failure also seemed to occur much more commonly when individuals were using assistive devices, such as a walker or cane. Qualitatively, the inferred hip and knee movements appeared much smaller than the real ones, aligning with our quantitative results below. It is possible recent advances that make mesh regression more robust to occlusion would help (Kocabas et al., 2021), but the implementation has not yet been released. ProHMR (Kolotouros et al., 2021) appears to produce more accurate results when including optimization of the mesh to the keypoints, but because this takes several seconds per frame, it is impractical to run on our entire dataset.

Accuracy of Kinematic Trajectories For each trial, we compared the trajectories computed with VIBE (Kocabas et al., 2020), directly from the 3D joint locations (Liu et al., 2020), and from the gait transformer to the ground truth. Two trials from these approaches are shown in Figure 2. We quantified the accuracy with the root mean squared error (RMS), measured in degrees for the joint angles and meters for the foot position, and the correlation coefficient. From the pretrained algorithms, we found the errors were fairly high and correlations were quite low for angles and foot positions. In comparison, the transformer predicted much more accurate and less noisy trajectories for these measures (Table 2).

To determine if more accurate kinematics could be computed from the 3D joint locations than our geometric approach, we trained a four layer MLP (fully connected, 256 units, GeLU non-linearity, layer scaling, labeled as MLP in Table 2 and Fig. 2) in place of the transformer. This also served as a lesion study for temporal context the transformer provides. We found this strategy did improve the accuracy and reduce the noise, but not as much as full transformer.

Gait event detection accuracy Most gait parameters (Table 1) are defined over a cycle between two successive times the same foot goes down. Thus, we first describe the accuracy detecting these event times. As described in Section 3, the model outputs the quadrature encoded phase with respect to the four gait events for each frame, from which we compute the phase on each frame, $\hat{\phi}_i(t)$, with Eq. 1. Event times are defined as when the phase crossed zero (specifically time between the frames where the zero crossing occurred). To compute the error, we matched each ground truth event to the nearest zero crossing (as there are multiple gait cycles per trial but only one annotated with events) and measured the time difference. The median error was 25ms for foot down events and 27ms for foot up events, with the 90% percentile for all errors being 83ms. This is fairly comparable to the 10ms error reported in Kidziński et al. (2019) for detecting foot down and up events from motion capture data acquired at 120Hz. The accuracy of the MLP was much worse (median 60ms) despite additional application of a Kalman smoother (Rauch et al., 1965) to the noisy outputs. This

	Right Hip	Left Hip	Right Knee	Left Knee	Right Foot	Left Foot
VIBE rms	20.12	19.93	19.77	19.94	0.28	0.26
3D rms	18.72	18.98	21.96	21.76	0.19	0.19
MLP rms	12.18	12.44	14.94	14.95	0.12	0.12
Transformer rms	9.43	9.33	11.64	11.54	0.08	0.08
VIBE r	0.53	0.53	0.29	0.26	0.62	0.66
3D r	0.39	0.40	0.27	0.28	0.73	0.74
MLP r	0.82	0.80	0.65	0.65	0.84	0.82
Transformer r	0.93	0.93	0.84	0.83	0.96	0.96

Table 2: Top section: Median RMS errors for joint angles (in degrees) and feet positions (in meters) using either VIBE, 3D joint locations, 3D joint locations transformed by an MLP, or the transformer output. Bottom section: Median correlation coefficient (R) between kinematic parameters and ground truth.

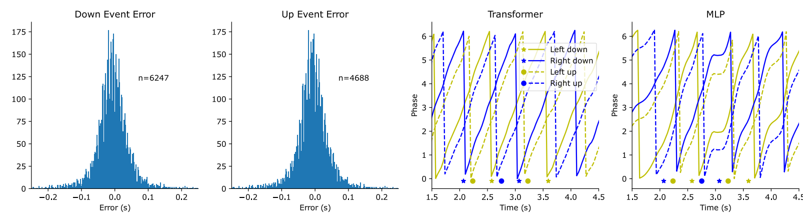

Figure 3: The left two panels show the error histograms for the down and up events, respectively. The third panel shows an example of the four gait phase outputs from the gait transformer with the ground truth events plotted, showing the tight alignment between the zero crossings and true events. The fourth panel shows the same thing for the MLP model after additional Kalman smoothing.

is consistent with our expectation that the transformer effectively combines information over the time series. The attention weights also suggest the transformer combines information over multiple gait cycles and displays sharp transitions between gait phases (Supplementary Fig. 7).

Cadence and Gait Phase Duration Some of the more common summary statistics computed for gait are the cadence (steps per minute, with two steps taken per gait cycle), time spent in single leg support on the left and right, and time spent in double limb support. We computed these over single gait cycles from the detected event times and found a close correspondence with ground truth values, as shown by the Bland Altman plots in Fig. 4. Multiple trials are collected for a given condition (e.g. with bracing versus without) and, in clinical practice, typically the average value is reported. The bottom row of Fig. 4 shows the tight correlation between these reported values and the average parameters measured by our approach.

Spatiotemporal Parameters Two important spatiotemporal gait parameters are step length and walking velocity. Step length is defined as the distance between the left and right foot when both feet are on the ground, which we extract from the difference between the foot kinematic traces at the time point each foot goes down (see diagram in Fig. 1). Walking velocity is the stride length (distance between two successive foot contacts) divided by the time between these events (Richards et al., 2012). We found computing the average pelvis velocity output over a gait cycle produced a more accurate walking velocity estimate than the product of the step length and the cadence, so report that. Fig. 4 shows these estimates are fairly accurate, but also reveals a slight systematic error. We suspect the output is biased towards the average in the face of uncertain information, and this systematic error could be calibrated out *post hoc*. We also lesioned the height token, and as expected found this worsened the accuracy for estimating both of these parameters, particularly step length.

Generalization To test whether our pipeline generalizes to views outside the training data (e.g. from the side), we tested it on Human 3.6M (Ionescu et al., 2014), which includes motion capture with multiple camera views. It is not designed for gait analysis, so does not have gait parameters or

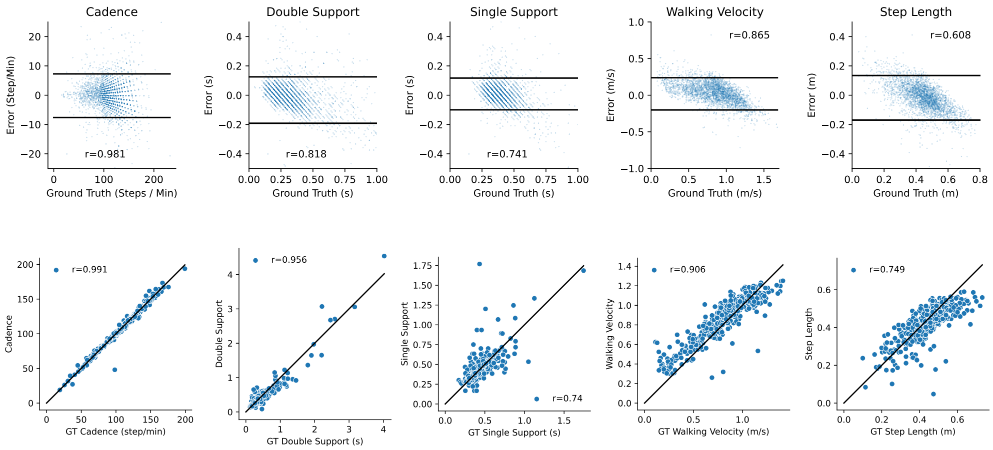

Figure 4: Top row: Bland-Altman plots of error for cadence, double support time and single support time with correlation coefficients. Horizontal bars are the 5% and 95% percentiles for error. Bottom row: average parameter values over trials with the same condition (e.g. bracing versus none).

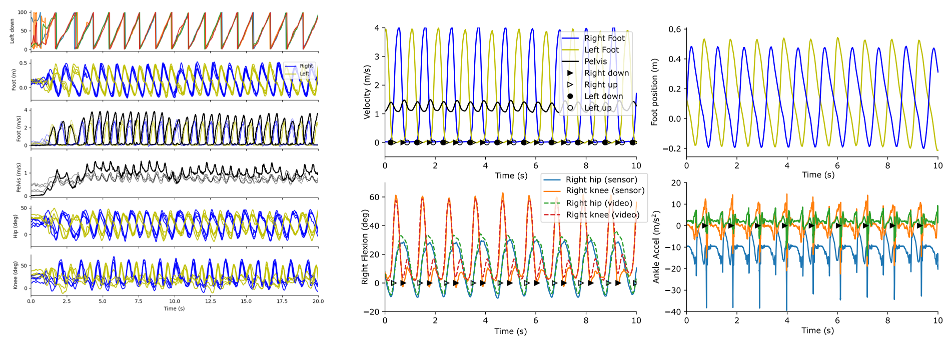

Figure 5: Left: Comparing our approach on four simultaneous views from Human 3.6 shows all views produce consistent estimates. Darker traces for foot and pelvis velocity show ground truth. Right: Results from our approach in clinic setting. The first row shows the model velocity and foot position outputs with the event times annotated. The second row compares these the joint angles to those estimated from wearable sensors, and also the detected event times compared to the accelerometer value at the ankle.

annotate foot up and down events. However, step events are apparent from the foot velocity and the walking velocity can be determined from the pelvis velocities. Fig. 5 shows samples of these traces from all four views, showing how similar the inferences are from all perspectives and the ground truth[2]. Lesioning the rotational augmentation worsened the generalization over views.

We also tested our pipeline in our intended context – a clinic visit. This differs from the training data in several ways. It is portrait video acquired at 1080x1920 on a smartphone over many gait cycles while walking with the subject down a hallway. The clinic hallways also appear different than the gait laboratory. We recorded video of a subject (in this case, an amputee using a prosthetic) walking down a hall. Ground truth was obtained from a wearable sensor system[3] on the prosthetic limb, which allows estimating the hip and knee angle and detecting foot down events from the accelerometer. The joint angles and event times from our pipeline closely matched the wearable system (Fig. 5, Right). Walking velocity was measured over 10 meters as 1.17 m/s, with our pipeline estimating 1.23 m/s .

[2] Video overlay from all views available at here and as supplemental video

[3] Reference removed during anonymous review

6 DISCUSSION

Our results show that 3D joint locations lifted from 2D keypoints contain enough information to characterize gait. However, at least on our gait lab dataset with a clinical population, the 3D joint locations themselves are not precise enough and required an additional task-specific transformer trained to output smooth sagittal plane kinematics and the timing of foot up and down events. From these interpretable features, we extracted gait parameters on a cycle-by-cycle basis and found a high correlation for cadence, double stance time, single support time, walking velocity and step length.

Compared to Kidziński et al. (2020), who predicted average gait parameters from 2D keypoint trajectories, our model outputs explainable features and allows a cycle-by-cycle analysis. This enables more precise characterization of gait variability in the clinic or home, which is associated with balance and quality of gait and possibly fall risk (Hausdorff et al., 2001; Hausdorff, 2005; Park et al., 2021). In addition, our correlation coefficients for walking speed and cadence (both near 0.9) exceed theirs (0.73 and 0.79, respectively) and we accurately estimate additional parameters.

Our results raise the question why the 3D keypoints did not yield more accurate kinematics. Due to the time required to process the thousands of videos, we could not exhaustively explore all possible components in our video processing pipeline. Each component (Zhang et al., 2020b; Wojke & Bewley, 2018; Sun et al., 2019; Zhang et al., 2020a; Liu et al., 2020) was selected for their competitive performance and promising results on initial testing, and our results will only improve with advances in human pose estimation, such as newer lifting algorithms (Shan et al., 2021; Gong et al., 2021). Despite this, the utility of the lifted 3D joints for gait analysis was limited. We speculate multiple factors contributed to this: (1) Individuals undergoing gait analysis often walk differently than would be seen in the public datasets, and algorithms trained on these might not generalized well to people with disabilities (Trewin et al., 2019). This is supported by the qualitative behavior we described above. Systematically evaluating the performance of all of the pipeline components on patient populations and people with disabilities is an important task for future work. (2) Optimizing lifting algorithms for 3D location error may not have an inductive bias towards biomechanically consistent results. (3) These algorithms are optimized for general HPE rather than specifically for gait analysis. Although, the graph attention spation-temporal convolutional network we used reports state of the art performance on lifting for walking accuracy (Liu et al., 2020). (4) Video resolution was 480x720 resolution and subjects sometimes spanned a small area, so performance might improve with different video conditions.

Ultimately, it would be preferable to have methods that produce both accurate 3D joint estimates and clinically useful kinematics, rather than training an additional component to accomplish this. Several recent studies incorporate either physics-based modeling into motion inference (Yuan et al., 2021; Shimada et al., 2021) or use a latent action space to constrain transitions to plausible human movements (Rempe et al., 2021), and additionally model foot contact events. We are excited that these approaches may produce more biomechanically accurate inferences, allow joint torque estimates, and we hope implementations of these will become available soon. Again, their generalization to clinical populations will need to be assessed. Another avenue is fine-tuning pretrained algorithms – either 3D lifting or physics-based – on the diverse set of gaits present in our dataset. This could include a loss function to ensure the representation is clinically interpretable and useful, but would also require careful design and possibly augmentation to ensure it generalizes outside the frontal views. Finally, self-supervised training of the gait transformer on additional, longer, unannotated samples of gait will likely further improve the performance when fine tuned to output gait kinematics.

7 CONCLUSIONS

Our approach improves the accuracy of kinematic trajectories estimated from monocular video and accurately predicts a number of gait parameters, but there is still a large gap between the precise biomechanical measurements made in a gait laboratory and what can be obtained from video. The utility of our approach will depend on the clinical question at hand. For measuring the parameters we explored and monitoring changes at home or during a clinical encounter, we are optimistic this method can be useful. Despite this, we hope this work primarily highlights the need for HPE to produce more clinically relevant and calibrated outputs.

REFERENCES

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.

G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020a.

MMTracking Contributors. MMTracking: OpenMMLab video perception toolbox and benchmark. `https://github.com/open-mmlab/mmtracking`, 2020b.

Kehong Gong, Jianfeng Zhang, and Jiashi Feng. Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8575–8584, June 2021.

J. M. Hausdorff. Gait variability: methods, modeling and meaning. *J Neuroeng Rehabil*, 2:19, Jul 2005.

J. M. Hausdorff, D. A. Rios, and H. K. Edelberg. Gait variability and fall risk in community-living older adults: a 1-year prospective study. *Arch Phys Med Rehabil*, 82(8):1050–1056, Aug 2001.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2016.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

M. P. Kadaba, H. K. Ramakrishnan, and M. E. Wootten. Measurement of lower extremity kinematics during level walking. *J Orthop Res*, 8(3):383–392, May 1990.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018.

R. M. Kanko, E. K. Laende, E. M. Davis, W. S. Selbie, and K. J. Deluzio. Concurrent assessment of gait kinematics using marker-based and markerless motion capture. *J Biomech*, 127:110665, Aug 2021a.

R. M. Kanko, E. K. Laende, G. Strutzenberger, M. Brown, W. S. Selbie, V. DePaul, S. H. Scott, and K. J. Deluzio. Assessment of spatiotemporal gait parameters using a deep learning algorithm-based markerless motion capture system. *J Biomech*, 122:110414, 06 2021b.

P. Khera and N. Kumar. Role of machine learning in gait analysis: a review. *J Med Eng Technol*, 44 (8):441–467, Nov 2020.

Ł. Kidziński, S. Delp, and M. Schwartz. Automatic real-time gait event detection in children using deep neural networks. *PLoS One*, 14(1):e0211466, 2019.

Ł. Kidziński, B. Yang, J. L. Hicks, A. Rajagopal, S. L. Delp, and M. H. Schwartz. Deep neural networks enable quantitative movement analysis using single-camera videos. *Nat Commun*, 11 (1):4054, 08 2020.

Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, and Carol Willing. Jupyter notebooks – a publishing format for reproducible computational workflows. In F. Loizides and B. Schmidt (eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pp. 87 – 90. IOS Press, 2016.

Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proceedings International Conference on Computer Vision (ICCV)*. IEEE, October 2021.

Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing.

Junfa Liu, Juan Rojas, Zhijun Liang, Yihui Li, and Yisheng Guan. A graph attention spatio-temporal convolutional network for 3d human pose estimation in video, 2020.

Wu Liu, Qian Bao, Yu Sun, and Tao Mei. Recent advances in monocular 2d and 3d human pose estimation: A deep learning perspective, 2021.

Adrian Llopart. Liftformer: 3d human pose estimation using attention models, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017.

S. Mehdizadeh, H. Nabavi, A. Sabo, T. Arora, A. Iaboni, and B. Taati. Concurrent validity of human pose tracking in video for measuring gait parameters in older adults: a preliminary analysis with multiple trackers, viewing angles, and walking directions. *J Neuroeng Rehabil*, 18(1):139, Sep 2021.

J. H. Park, H. Lee, J. S. Cho, I. Kim, J. Lee, and S. H. Jang. Effects of knee osteoarthritis severity on inter-joint coordination and gait variability as measured by hip-knee cyclograms. *Sci Rep*, 11 (1):1789, 01 2021.

Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

H. E. Rauch, C. T. Striebel, and F. Tung. Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3(8):1445–1450, August 1965. doi: 10.2514/3.3166.

Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. *CoRR*, abs/2105.04668, 2021. URL `https://arxiv.org/abs/2105.04668`.

Jim Richards, David Levine, and Michael Whittle. *Whittle's Gait Analysis*. Elsevier, 2012.

Nidhi Seethapathi, Shaofei Wang, Rachit Saluja, Gunnar Blohm, and Konrad P. Kording. Movement science needs different pose tracking algorithms, 2019.

Wenkang Shan, Haopeng Lu, Shanshe Wang, Xinfeng Zhang, and Wen Gao. Improving robustness and accuracy via relative information encoding in 3d human pose estimation. *CoRR*, abs/2107.13994, 2021. URL https://arxiv.org/abs/2107.13994.

Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *arXiv preprint arXiv:2006.12075*, 2020.

Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics*, 39(6), dec 2020.

Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. Neural monocular 3d human motion capture with physical awareness. *ACM Trans. Graph.*, 40(4), July 2021. ISSN 0730-0301. doi: 10.1145/3450626.3459825. URL https://doi.org/10.1145/3450626.3459825.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer, 2021.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers, 2021.

Shari Trewin, Sara Basson, Michael Muller, Stacy Branham, Jutta Treviranus, Daniel Gruen, Daniel Hebert, Natalia Lyckowski, and Erich Manser. Considerations for ai fairness for people with disabilities. *AI Matters*, 5(3):40–63, December 2019. doi: 10.1145/3362077.3362086.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

J. Verghese, A. LeValley, C. B. Hall, M. J. Katz, A. F. Ambrose, and R. B. Lipton. Epidemiology of gait disorders in community-residing older adults. *J Am Geriatr Soc*, 54(2):255–261, Feb 2006.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2019.

Nicolai Wojke and Alex Bewley. Deep cosine metric learning for person re-identification. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 748–756. IEEE, 2018. doi: 10.1109/WACV.2018.00087.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3645–3649. IEEE, 2017. doi: 10.1109/ICIP.2017.8296962.

World Health Organization. Falls, 2021. URL https://www.who.int/news-room/fact-sheets/detail/falls.

G. Wu, S. Siegler, P. Allard, C. Kirtley, A. Leardini, D. Rosenbaum, M. Whittle, D. D. D'Lima, L. Cristofolini, H. Witte, O. Schmid, and I. Stokes. ISB recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion–part I: ankle, hip, and spine. International Society of Biomechanics. *J Biomech*, 35(4):543–548, Apr 2002.

G. Wu, F. C. van der Helm, H. E. Veeger, M. Makhsous, P. Van Roy, C. Anglin, J. Nagels, A. R. Karduna, K. McQuade, X. Wang, F. W. Werner, and B. Buchholz. ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion–Part II: shoulder, elbow, wrist and hand. *J Biomech*, 38(5):981–992, May 2005.

Dimitri Yatsenko, Jacob Reimer, Alexander S. Ecker, Edgar Y. Walker, Fabian Sinz, Philipp Berens, Andreas Hoenselaar, R. James Cotton, Athanassios S. Siapas, and Andreas S. Tolias. Datajoint: managing big scientific data using matlab or python. *bioRxiv*, 2015. doi: 10.1101/031658. URL `https://www.biorxiv.org/content/early/2015/11/14/031658`.

Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. Simpoe: Simulated character control for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7159–7169, June 2021.

Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020a.

Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking, 2020b.

Ce Zheng, Wenhan Wu, Taojiannan Yang, Sijie Zhu, Chen Chen, Ruixu Liu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey, 2020.

A Appendix

A.1 Video Processing

DataJoint Pipeline To manage these dependent steps from thousands of videos, we implemented a human pose estimation pipeline using DataJoint (Yatsenko et al., 2015), which handles both the data management and computational dependencies with a database backend. Figure 6 shows the computational dependencies for this analysis, with each node reflecting a separate MySQL table linked via foreign primary keys.

Wrappers for algorithms implementing each of these steps were implemented in our DataJoint Pose Pipeline using a common data format in the database, making it straightforward to the pipeline to change the algorithm for each step. We used the official implementation and released weights for each algorithm.

Tracking and Annotation Videos typically included multiple people including staff helping with data acquisition and sometimes physical therapists providing assistance with gait, so the first step was identifying the person of interest. We do this by first running a pretrained algorithm to produce bounding boxes with subject IDs, and then manually annotating which subject ID corresponds to the person of interest. In some videos multiple bounding boxes would be detected for the individual in different parts of the video (e.g. fragmentation), in which downstream algorithms took the unions of these bounding boxes. In videos where either the person was not detected in a large number of frames or critically if a bounding box switched from tracking the person of interest to another person (most commonly a nearby physical therapy), the video was marked as invalid. For these invalid videos, we ran alternative tracking algorithms which would frequently allow reliable tracking. The manual annotation of these thousands of videos was facilitated with a custom tool using PyWidgets in a Jupyter notebook (Kluyver et al., 2016) that would query the database, show the next unannotated video, and insert the annotation into the database after clicking the appropriate track. In videos processed with multiple tracking algorithms that were annotated as valid, we used the one with the highest fraction of frames where the person was detected for subsequent processing.

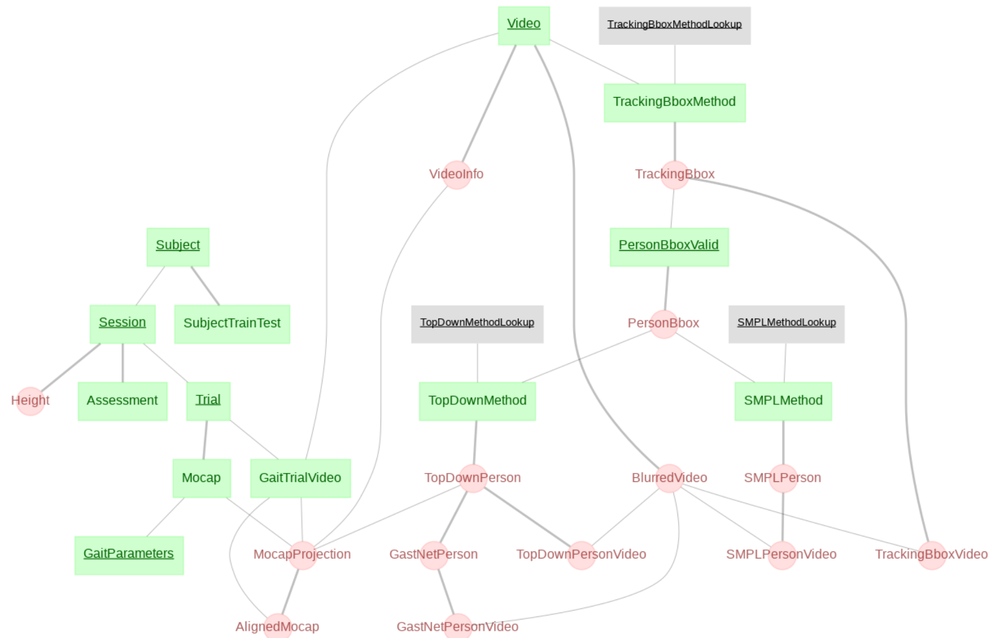

Figure 6: DataJoint schema for computational pipeline. The right side shows the general video processing pipeline and the left side the organizational structure specific to the gait lab data. Lookup tables allow selecting different algorithms for video processing component. Circular nodes are computed can be computed once requisite parent nodes exist.

The majority of our videos were successively annotated with FairMOT (n=5358, Zhang et al. (2020b)). The next most common tracking algorithm we used was DeepSORT (n=2891, Wojke & Bewley (2018); Wojke et al. (2017)). A few videos were also processed with TransTrack (n=164) (Sun et al. (2021)) and MMTrack (n=55, Contributors (2020b)).

2D keypoint and 3D joint locations Two-dimensional joint locations are extracted with an HRNet (Sun et al., 2019; Wang et al., 2019) trained with a Distribution-Aware Coordinate Representation of Keypoints (Zhang et al., 2020a) on the COCO dataset (Lin et al., 2014). We use the implementation and pretrained weights from the MMPose package (Contributors, 2020a), which has the additional benefit of providing a wide range of network architectures and training strategies using a consistent API. The 2D keypoints sequences are lifted to 3D coordinates (relative to the pelvis) using a graph attention spation-temporal convolutional network (GAST-Net) (Liu et al., 2020) with the official implementation and pretrained weights and a receptive field of 27 frames. As with the tracking algorithms, Pose Pipeline wrappers for these algorithms directly pull the videos and bounding boxes from DataJoint and the results are inserted directly into the database.

Computing the 3D keypoints took approximately a month using an Nvidia A100 and two 3090RTX GPUs and two of the authors performing manual annotation for several hours a day.

Reprojection error Although the synchronization between the motion capture system and the video camera is hardware triggered, there is still typically 100-200 ms of offset. We correct this by jointly optimizing for the intrinsic and extrinsic camera parameters as well as a temporal offset that minimizes the reprojection error of the hip, knee and ankle joints into the image plane. The camera properties are first initialized with the OpenCV calibrateCamera method (Bradski, 2000), using the the sequence of paired 2D keypoint locations from the video and the ground truth 3D coordinates from the motion capture data. This initialization assumes zero time offset and uses the time range with both high keypoint confidences and where the motion capture markers were visible. A differentiable reprojection loss is then computed with respect to the camera parameters and the temporal offset by first computing the 3D joint location with the time offset using linear interpolation, projecting these through a simple camera model (i.e. no distortion parameters), and then finally computing the Huber loss between these projected coordinates and the detected joint locations in the image. This loss is optimized with Jax (Bradbury et al., 2018).

Trials were only included for analysis if the average reprojection Huber loss across the six joints in the leg was less than 10 pixels and the absolute value of the offset was less than 200ms. Only 219 out of 8468 trials were excluded for this reason, confirming that 2D keypoint detection generally performed well on rehabilitation subjects.

After screening the data for poor bounding box annotations, reprojection errors, trials with insufficient frames that the person was visible, and trials with an invalid sequence of gait events our final dataset contained 6747 trials for training and 1592 trials for testing.

A.2 Attentional Weights

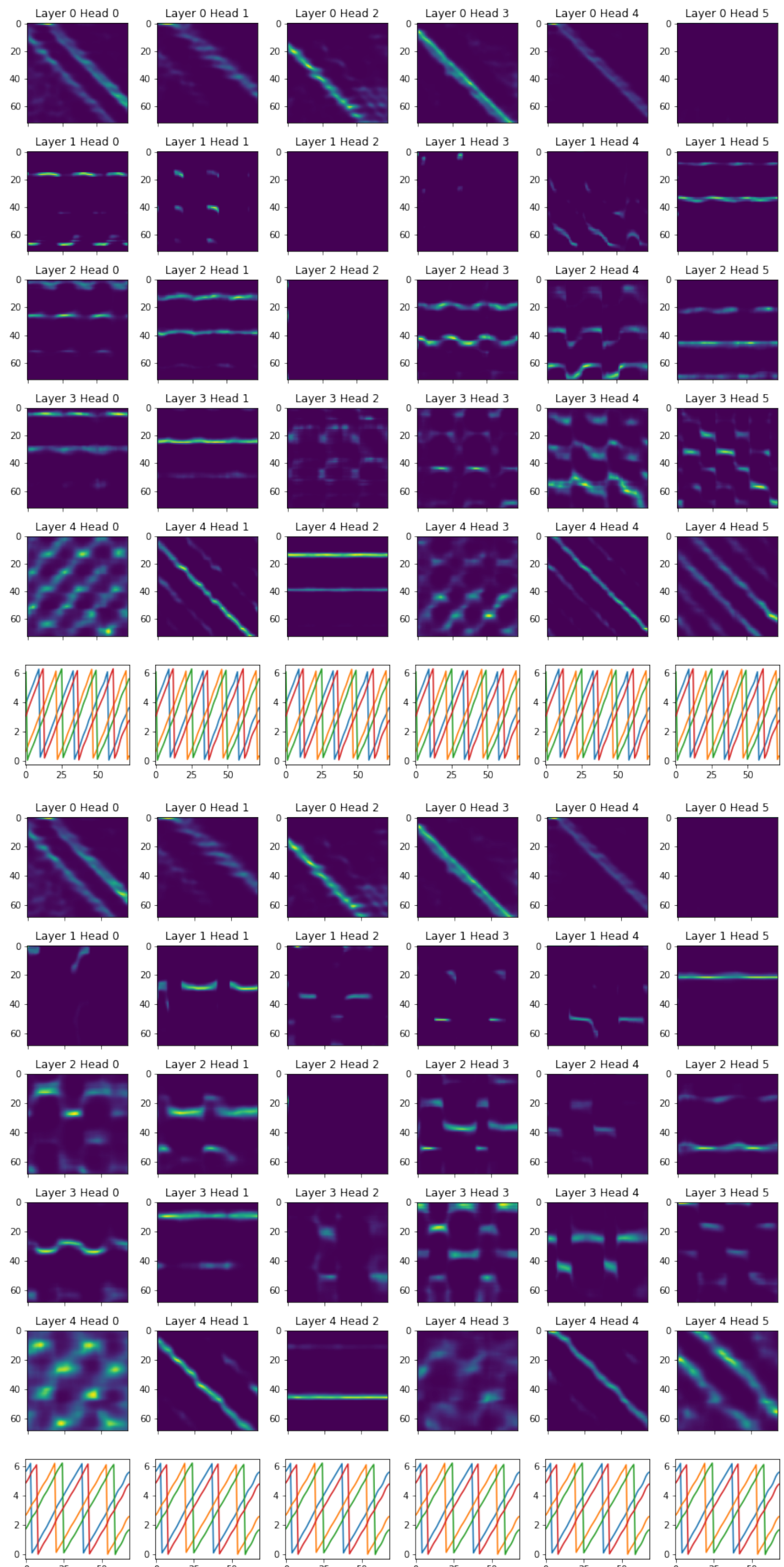

Figure 7: Two examples of the attention weights throughout the transformer (all layers and heads are shown), with the phase of the four gait events at the bottom. The attention scores were transpose to align with the gait phase traces below. The attention pattern period matches the gait period, and even develops sharp transitions aligned with foot events despite the quadrature encoded phases being smooth.