# GRF_MV:Ground Reaction Force Estimation from Monocular Video

Juni Katsu (2246810)
jxk010@student,bham.ac.uk
BSc Computer Sci w AI, University of Birmingham
Birmingham, United Kingdom

Esha Daspupta
esha.dasgupta@gmail.com
University of Birmingham
Birmingham, United Kingdom

Hyung Jin Chang
h.j.chang@bham.ac.uk
University of Birmingham
Birmingham, United Kingdom

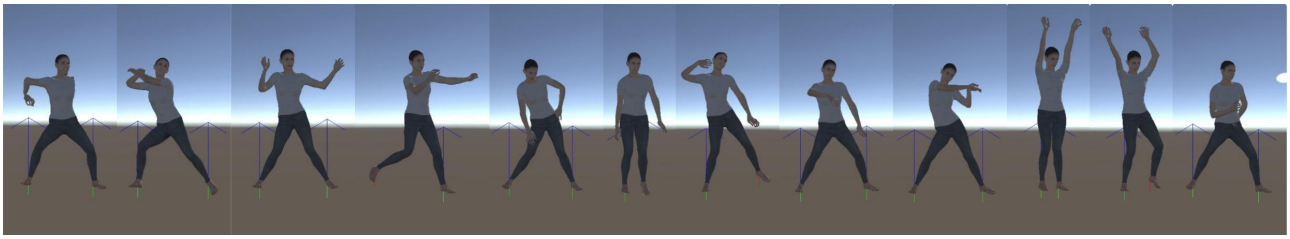


**Figure 1: Sequences of estimated Ground Reaction Force from a dancing video**

## ABSTRACT

Estimating ground reaction forces (GRFs) from monocular video has essential applications in biomechanics, sports performance analysis, injury prevention and rehabilitation. However, it is a challenging problem due to the complexity of human motion, the limited availability of training data, and the difficulty of estimating contact and forces from 2D video alone. This paper presents a novel approach for estimating ground reaction forces (GRFs) from monocular video by combining deep learning-based 3D human mesh recovery with physics-based optimization. Existing techniques for measuring GRFs rely on specialized sensors or multiple camera setups, limiting their applicability outside lab settings. In contrast, the proposed approach requires only a single video camera, making it suitable for deployment in sports, clinical and home environments. A deep neural network is trained to recover 3D human mesh parameters from each frame, which are further refined using physics-based optimization. GRFs are then estimated from the 3D foot velocities and contact modelling. The approach is evaluated on the GroundLink dataset, demonstrating improved accuracy over prior methods. However, further work is needed to improve generalization performance, computational efficiency, and physical plausibility. The code and trained models are available for use by the research community.

## CCS CONCEPTS

• **Computing methodologies** → **Image and video acquisition**.

## KEYWORDS

Ground Reaction Force, SMPL-X, Monocular Video, Inverse Kinematics, Physics-Based Method

Authors' addresses: Juni Katsu (2246810), jxk010@student,bham.ac.uk, BSc Computer Sci w AI, University of Birmingham, School of Computer Science, Birmingham, United Kingdom, B15 2TT; Esha Daspupta, esha.dasgupta@gmail.com, University of Birmingham, Birmingham, United Kingdom; Hyung Jin Chang, h.j.chang@bham.ac.uk, University of Birmingham, Birmingham, United Kingdom.

## 1 INTRODUCTION

Measuring and analyzing human motion and contact forces offer valuable insights for a wide range of applications such as sports biomechanics, injury prevention, rehabilitation, and human-computer interaction. One of the key approaches is to measure Ground Reaction Force (GRF). GRFs refer to the force exerted by the ground on a body in contact with it. Conventionally, force plates or wearable pressure sensors are used to record it in controlled laboratory settings. These approaches are expensive and intrusive and restrict the range and naturalness of motions that can be captured. Some deep learning models are trained to regress GRFs using measured forces. Physics-based optimization can also be used to constrain the estimated forces and body kinematics to be physically plausible.

With rapid progress in computer vision, multiple deep learning networks are established to capture human body motions without predefined markers, recovering to 3D body mesh from 2D video frames taken with single (monocular) or multiple cameras. Parametric body models like SMPL and SMPL-X represent the 3D geometry and pose of the human body. However, recovering 3D body shape, pose, and contact forces from 2D projected images is extremely challenging due to depth ambiguity, occlusions, variations in body shape and clothing, and the complexity of human motion.

This paper tackles the problem of estimating GRFs from monocular RGB video by leveraging recent advances in 3D human mesh recovery and physics-based optimization with inverse kinematics approaches. We introduce a new framework for video-based GRF estimation using monocular video as the only input, combined with simple physics-based inverse kinematics optimization, including the new factor to deep learning neural network-based optimization. First, a parametric 3D body model (SMPL-X) is fit to each frame of a monocular video using a proposed deep neural network, providing an initial estimate of the body's pose and shape. Next, the motion of the body and feet are refined using physics-based optimization

to ensure they remain in contact with the ground plane. Finally, the GRFs are calculated from the optimized foot motions and evaluated with a data-driven model trained on the GroundLink dataset.

We focus on single-person dynamic motion from single camera token video of dancing. Our approach matches current state-of-the-art methods on 3D mesh recovery and estimating GRFs with improvement in their complexity, requiring less computational resources compared to existing methods. We use Unity to demonstrate our estimated Ground Reaction Force on character animation by modifying generated keypoints for each frame of a single video. To evaluate our method, we evaluate the accuracy of calculated GRF on captured motion and GRFs from the data-driven GRF dataset Groundlink because we lack of actual GRF data on a dancing video.

## 2 RELATED WORK

Estimating human motion and contact forces from videos have been a longstanding challenge in computer vision and graphics. This section reviews the relevant background material and prior work on 3D human mesh recovery, physics-based optimization, and data-driven GRF prediction.

### 2.1 3D mesh recovery

3D mesh recovery from videos has been an emerging reserach topic in the computer vision community with many applications. While recovering the full 3D human shape and pose from a single image is a fundamentally ill-posed problem, many previous works have contributed to the improvement of its accuracy using different deep-learning neural networks in whole-body pose estimation from videos. Parametric 3D body models like SMPL[Loper et al. 2015], and SMPL-X[Pavlakos et al. 2019] are a compact representation that can be predicted using deep learning neural networks. Methods like HMR[Kanazawa et al. 2017] and SPIN[Kolotouros et al. 2019] use CNNs to regress body model parameters from pixels of images directly. Recent approaches estimate 2D-keypoints of the human body in the image and silhouettes as intermediate representations to improve accuracy. Combining 2D keypoint estimation with temporal information inferred from recurrent networks (VIBE)[Kocabas et al. 2019], or motion discriminator (TCMR)[Choi et al. 2020] or transformers (MeshTransformer)[Lin et al. 2020] as examples, shows significant improvement in the accuracy of the captured motion. However, most of these methods are trained on datasets of posed subjects and do not handle foot contacts or the depth of the camera of the environment.

Another work of interest is HybrIK[Li et al. 2020], which fits the SMPL body model to each 2D-keypoint estimated frame of video using the HR network [Sun et al. 2019] and inverse kinematics(IK) optimization to enhance joint angles and twists, which potentially reduce foot sliding. HybrIK-X[Li et al. 2023] extends this to handle face's and hand's expression. In this paper we build upon HybrIK-X[Li et al. 2023] for robustly fitting the SMPL-X body model meshes to videos as the input to our GRF prediction pipeline.

### 2.2 Ground Reaction Force Estimation

Ground reaction force (GRF) is the force exerted by the ground on a body in contact based on factors such as weight, force of movement, terrain friction, environmental forces and so on. GRF provides valuable insights into human motion, balance, and the distribution of force along the musculoskeletal system, making it essential for various biomedical applications such as sports performance analysis, injury prevention, and rehabilitation. However, accurately measuring GRF remains a challenging task. Traditionally, GRF is measured using specialized equipment like force plates[Group 2020] or pressure-sensitive insoles[Shahabpoor and Pavic 2017][Ancillao et al. 2018].

Currently, several datasets have been developed for the task to enhance the accuracy of GRf estimation. In the field of biomechanics and biomedical engineering, a few new datasets of motion capture with GRFs [Kulbacki et al. 2021] and [Zhu et al. 2023] datasets are available. However, these datasets are often limited to specific aspects of movement, target stages and diseases in pathological subjects. Datasets potentially suitable for our approach are UnderPressure[Mourot et al. 2022] and PSU-TMM100[Yang et al. 2021]. The UnderPressure dataset contains synchronized video, motion capture, and pressure insole data for various activities, providing detailed foot pressure maps and GRF data, but lacks 3D body shape information. The PSU-TMM100 dataset includes video, motion capture, and force plate data for 100 subjects performing various activities, but the 3D body shape is represented using a simplified marker-based model.

The most suitable dataset for our approach is GroundLink[Han et al. 2023], which provides synchronized video, motion capture, and force plate data for 60 subjects performing a wide range of activities. Notably, the motion capture data in GroundLink is visualized using the SMPL-X body model, and all of the body joints are adjusted to align with its model, enabling smooth conversion from recovered 3D mesh body from videos. Furthermore, force plate data from GroundLink can be used as ground truth for model evaluation.

### 2.3 Physics-based Character Animation

Accurately estimating foot contact and ground reaction forces from monocular video is a challenging task due to the inherent ambiguity in 2D images and the complex dynamics of human motion. To address this issue, researchers have explored physics-based optimization techniques that incorporate biomechanical constraints and principles to improve the plausibility and accuracy of the estimated forces. Rempe et al. [Rempe et al. 2020] proposed a method that combines a deep neural network for 3D human motion estimation with a physics-based optimization module for foot contact optimization. The neural network predicts the 3D joint positions and velocities from the monocular video, while the optimization module refines the foot contacts and estimates the ground reaction forces based on the predicted motion and a simplified physics model. By alternating between the neural network and optimization modules, they achieve more accurate and physically plausible results.

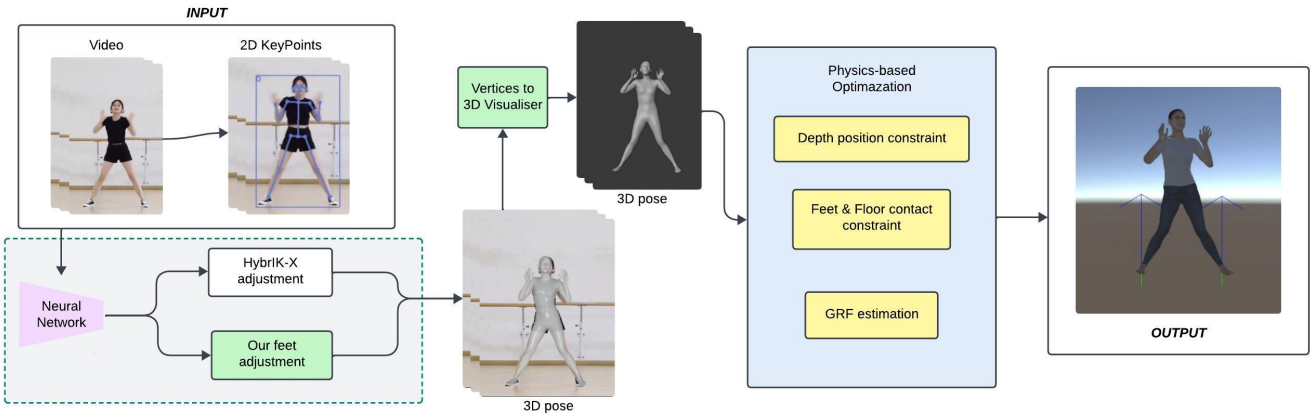



**Figure 2: pipeline Overview**

Recent works have further advanced physics-based optimization techniques for human motion estimation and synthesis from videos. Shimada et al. [Xie et al. 2021] introduced a physics-based method that estimates human motion from videos by optimizing the joint torques and contact forces to minimize the discrepancy between the observed and simulated motion. They incorporate biomechanical constraints and a learned motion prior to improve the accuracy and realism of the estimated motion. Yi et al. [Yi et al. 2022] proposed the Physical Inertial Poser (PIP), a real-time human motion tracking system that combines sparse inertial sensor data with physics-based optimization. PIP optimizes the joint angles and velocities to match the sensor measurements while satisfying physical constraints such as joint limits and contact dynamics.

Another promising direction is the integration of physics-based optimization with deep learning techniques. Yuan et al. [Yuan et al. 2023] introduced PhysDiff, a physics-guided human motion diffusion model that generates physically plausible human motions by incorporating a physics-based optimization step into the diffusion process. PhysDiff ensures that the generated motions satisfy physical constraints such as joint limits, contact dynamics, and balance. Similarly, Tripathi et al. [Tripathi et al. 2023] proposed a method for 3D human pose estimation that leverages intuitive physics to improve the accuracy and plausibility of the estimated poses. They use a physics-based optimization module to refine the initial pose estimates obtained from a deep neural network, taking into account physical constraints such as balance and contact dynamics.

# 3 METHODOLOGY

## 3.1 Overview

This section describes our framework for estimating GRFs from monocular video, summarised in Figure 2. Given an input video, human poses are estimated for each frame in 3D meshes using HybrIk-X[Li et al. 2023]. We connect the sequences of joint coordinates of the 3D human meshes to generate an animation. The animation is then refined in the physics-based engine to optimize its foot contact and kinematic plausibility. Finally, we apply the GRF equation to the animated model to calculate and evaluate GRFs.

## 3.2 3D Mesh Recovery

For our framework, we have chosen HybrIK-X, a state-of-the-art model for 3D mesh recovery, from amongst the existing models available. HybrIK-X is trained on the SMPL-X mesh dataset, which is the full-body extension of the SMPL mesh recovery model. SMPL-X allows the use of shape, expression, and pose parameters, enabling the recovery of a more comprehensive and detailed 3D human mesh. SMPL-X is a relatively new mesh model and has gained attention in the research community. However, the number of available models and datasets for SMPL-X is still limited.

One of the critical advantages of HybrIK-X is its independence from other applications, making it easy to implement and integrate into our framework. Additionally, the clear relationship between each dependency in HybrIK-X facilitates modifications and adaptations to suit our specific requirements. This flexibility is particularly valuable in our research, as it allows us to fine-tune and optimize the model for our specific use case. Furthermore, despite its ease of use and adaptability, HybrIK-X delivers relatively good results in terms of 3D mesh recovery accuracy, making it a suitable choice for our framework.

*3.2.1* ***HybrIK-X.*** HybrIK-X has enhanced twist and swing of joint rotations by recursively going along the kinematic tree. The original rotations of twist and swing and corresponding networks are proposed. To calculate swing rotation, HybrIK-X utilizes a combination of neural networks and 3D joint information. Separate neural networks are employed to estimate a low degree of freedom twist, further refining the joint rotations. The proposed equation for this enhancement is as follows: Given the start template body part vector $\vec{t}$ and the target vector $\vec{p}$.

$$R = \mathcal{D}(\vec{p}, \vec{t}, \phi) = \mathcal{D}_{sw}(\vec{p}, \vec{t})\mathcal{D}_{tw}(\vec{t}, \phi) = R_{sw}, R_{tw} \quad (1)$$

where $\phi$ is the twist angle estimated by a neural network, $\mathcal{D}_{sw}$ is a closed-form solution of the swing rotation and $\mathcal{D}_{tw}$calculates twist rotation from $\phi$. $R$ here should satisfy the condition $\vec{p} = R\vec{t}$. This equation ensures that the swing rotation accurately aligns the template body part vector with the target vector, resulting in more natural and realistic joint movements in the recovered 3D mesh.

HybrIK-X further introduces the divide-and-conquer IK process. The whole-body kinematic tree is divided into four sub-trees and is processed individually. This results in a closer relation with the root joint for each joint and more robust to bone key points' noises. For more details about HybrIK-X, please refer to [Li et al. 2023].

*3.2.2* ***Optimization.*** HybrIK-X demonstrates impressive performance in estimating joint rotations and positions through its well-designed inverse kinematics strategy. However, despite these advancements, significant errors in feet positions can still be observed. Accurate contact estimation between the feet and the floor is crucial for the reliable estimation of ground reaction forces (GRFs). To address this challenge, we introduce an additional loss term specifically designed to improve the accuracy of foot-ground contact estimation.

Ideally, the most comprehensive approach would be to estimate the ground plane alongside the 3D human pose. We explored this possibility by leveraging the planeRCNN model proposed by [Liu et al. 2018], which predicts the plane and camera coordinates from images. Although planeRCNN generated multiple potential plane coordinates, we found that these estimates did not perfectly align with our requirements due to discrepancies in the predicted camera parameters between the two models. Furthermore, while other state-of-the-art plane detection methods offer high accuracy, they primarily focus on plane segmentation and lack the necessary coordinate information for our purposes.


**Figure 3: Offset between ground and ankle joint**

To overcome these limitations, we propose a practical alternative approach. Instead of estimating the ground plane directly, we manually select a frame from the video where the person is standing on the ground. We then calculate the average of the z-axis coordinates of the right and left ankle in that frame and set it as the ground truth $G_z$. This value serves as a reference for comparison with the current feet position $E_0$ along the z-axis. The loss function is designed to adapt based on the relationship between $F_z$ and $G_z$.

$$\mathcal{L}_{feet} = \begin{cases} E_0, & \text{if } F_z - G_z - offset < 0. \\ E_1, & \text{otherwise.} \end{cases} \tag{2}$$

Feet clipping occurs when the estimated feet positions penetrate the ground plane, resulting in visually implausible and physically incorrect poses. The loss function $E_0$ is designed to minimize the occurrence of feet clipping as much as possible. Equation 3 calculates the mean squared error(MSE) between the predicted feet positions $\alpha_k$ and the ground-truth feet positions $\hat{\alpha_k}$ for $K$ feet joints, which in our case is typically two (right and left ankle). This will encourage the predicted feet positions to align closely with the ground truth, effectively pushing the feet upwards to avoid clipping. The use of mean squared error is particularly effective in this scenario as it penalizes larger deviations more heavily, providing a strong incentive to eliminate feet clipping completely.

$$E_0 = \frac{1}{K}\sum_{k=1}^{K} \parallel \alpha_k - \hat{\alpha_k} \parallel^2 \tag{3}$$

However, setting the loss term to minimize the distance between the feet and the ground may result in another undesirable artefact: constant feet floating. In reality, the feet are not always in contact with the ground during various movements such as walking, running, or jumping. Therefore, we cannot rely on a loss term that consistently pushes the feet towards the ground plane. To address this challenge, we introduce the loss term $E_1$, which is applied when the feet are predicted to be above or on the ground plane ($F_z - G_z - offset \geq 0$).

$$E_1 = -\sum_{k=1}^{K} \omega_0 log(1 - \bar{\alpha}_k) \tag{4}$$

where $\bar{\alpha}_k = (\alpha_k - \hat{\alpha_k})/\sigma_k$ for $K$ feet joints and $\sigma_k$ denotes the standard deviation. Equation $E_1$ employs a log-likelihood loss to create a gentle loss curve that allows for the possibility of feet floating. The weight $\omega_0$ controls the steepness of the loss curve, allowing for adjustments based on the specific requirements of the dataset or application. Using a log-likelihood loss, we ensure that the penalty for feet floating gradually increases as the distance between the feet and the ground plane grows. This approach provides a balanced trade-off between keeping the feet close to the ground and allowing for natural floating positions during various movements.

**From pretrained model:** The HybrIK-X pretrained model contains three key components: pose, camera, and twist angle estimation. For pose estimation, a conventional RLE neural network is used, which leverages a fully connected layer to estimate the joint coordinates $p_k$ and their corresponding standard deviations $\sigma_k$. To prevent overfitting, both $p_k$ and $\sigma_k$ are represented as single-dimensional values. The loss function used to train the 3D keypoint estimation is formulated as a log-likelihood loss, where $Q$ represents the probability density function of the standard Laplace distribution, and $G$ is the distribution learned by the RLE network. The term $\bar{p_k}$ denotes the difference between the estimated and ground-truth joint positions. The loss to train the 3D keypoint is formulated as:

$$\mathcal{L}_{pose} = -\sum_{k=1}^{K} logQ(\bar{p}_k) - logG_\psi + 3log\sigma_k \tag{5}$$

For the task of camera position estimation, a weak-perspective camera model with a focal length of 1 meter is employed by HybrIK-X. Estimating the scale factor $S$ from a monocular RGB image is considered an ill-posed problem due to the inherent ambiguity in determining the absolute scale of the scene from a single view. This ambiguity arises because a larger object farther away from the camera can appear the same size as a smaller object closer to the camera, making it challenging to distinguish between the two scenarios without additional information. To address this issue, an iterative camera estimation method is adopted. This method regenerates the projection iteratively to minimize the projection error within the given constraints. The initial scale factor estimation $s^0$ is regressed by the neural network and supervised by the $l2$ loss:

$$\mathcal{L}_{cam} = \| s^0 - \hat{s} \|^2 \tag{6}$$

where $\hat{S}$ is the ground-truth scale factor.

As described in the HybrIK-X section, the twist and swing rotations of the estimated 3D poses are enhanced. While the swing rotation can be easily calculated as it is perpendicular to the start template body part vector $\bar{t}$ and the target vector $\bar{p}$ [Li et al. 2023], the twist angle requires a separate neural network for estimation. To estimate the scale value $\phi_k$ for calculating the twist angle, a 2-dimensional vector $(cos\phi_k, sin\phi_k)$ is chosen instead of direct regression to avoid the discontinuity problem. The loss function applied to train the twist angle estimation network is an $l2$ loss, as shown in the equation below, where $\hat{\phi_k}$ denotes the ground-truth twist angle for the $k$-th joint.

$$\mathcal{L}_{tw} = \frac{1}{K} \sum_{k=1}^{K} \| (cos\phi_k, sin\phi_k) - (cos\hat{\phi}_k, sin\hat{\phi}_k) \|^2 \tag{7}$$

**From SMPL-X model:** The SMPL-X additional parameters: shape $\beta$, expression $\psi$ and rotation $\rho$ are trained individually to obtain rest pose skeleton with additive offsets. $L2$ loss is calculated for each parameter:

$$\mathcal{L}_{shape} = \| \beta - \hat{\beta} \|^2 \tag{8}$$

$$\mathcal{L}_{exp} = \| \psi - \hat{\psi} \|^2 \tag{9}$$

$$\mathcal{L}_{rot} = \| \rho - \hat{\rho} \|^2 \tag{10}$$

where $\hat{\beta}$, $\hat{\psi}$ and $\hat{\rho}$ are the ground-truth for each parameter.

This is the full loss term for training proposed 3D pose estimation from monocular video method. It is formulated as:

$$\mathcal{L} = \mathcal{L}_{pose} + \mu_1 \mathcal{L}_{cam} + \mu_2 \mathcal{L}_{shape} + \mu_3 \mathcal{L}_{exp} + \mu_4 \mathcal{L}_{rot} + \mu_5 \mathcal{L}_{tw} + \mu_6 \mathcal{L}_{feet} \tag{11}$$

The network minimised the loss calculated to obtain the best estimation. $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$ and $\mu_6$ are weights of each loss term and will be learned during training.

## 3.3 Physics Based Optimization

In this section, we discuss the physics-based optimization techniques employed to enhance the stability and realism of the estimated 3D poses. The 3D poses obtained from section above for each image frame are connected to form a sequence of animation in Blender, which is then exported to Unity for further optimization. Unity is chosen as the platform for this task due to its accessibility, wide range of tools, and our previous experience with the software, making it easy to use and integrate into our pipeline.

*3.3.1* ***Feet constraint.*** One of the main issues addressed in this section is the instability of the feet in the existing work in the field of 3D human mesh estimation.. To tackle this problem, we perform feet adjustment in Unity. However, since the previous section does not include plane estimation due to the lack of existing models and the requirement for predicting ground reaction force, we manually set the plane in Unity.

Unity provides a feature called rigid body, which is a component that enables physical simulation of objects. Rigid bodies are affected by forces, collisions, and other physical interactions, making them essential for creating realistic dynamics in a virtual environment. By attaching a rigid body component to the feet meshes, we can simulate their interaction with the ground plane and ensure proper contact.

To accurately detect collisions between the feet and the ground, we utilize Unity's mesh collider component. A mesh collider is a type of collider that closely fits the shape of a 3D mesh, providing precise collision detection. By assigning mesh colliders to both the feet and the ground plane, we can detect when the feet come into contact with the ground and prevent them from passing through it.

However, the mesh collider alone provides only a soft constraint, meaning that it doesn't completely prevent feet clipping. To address this issue, we implement an additional vertical laser-based constraint. From each ankle and toe joint, two lasers are cast downwards with a length equal to an offset value. When these lasers hit the ground, it indicates that the feet are in contact with the floor. We then apply a position constraint to the ankle and toe joints, preventing them from moving lower than the plane.

The combination of the laser-calculated position constraint and the collision detector effectively prevents feet clipping. However, this approach only fixes the position of the feet joints, such as the ankle and toe, and does not address the positioning of the entire leg or potentially the full body. This limitation can lead to unpleasant deformations of the leg mesh and incorrect knee positions.

To resolve this issue, we employ momentary inverse kinematics (IK). Whenever the feet are about to clip through the floor, the hard constraint keeps them above the floor, and the corresponding leg's knee follows an inverse kinematics solution instead of its original animation. Inverse kinematics is a technique that determines the joint positions based on the end effector's position, which in our case is the ankle coordinates. Unity provides the capability to manually adjust the target position $(T)$ that the knee should follow.

To minimize the deviation from the original animation, the coordinates of the target position $(T)$ for the knee are calculated by following:

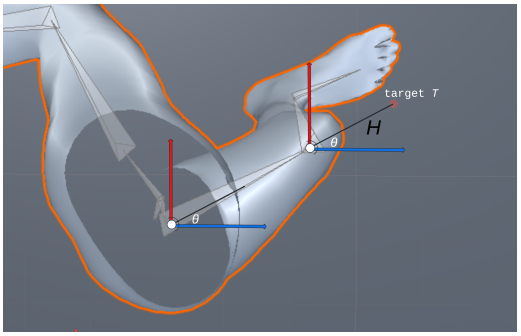



**Figure 4: Inverse kinematics target calculation**

$$T = (x_{knee} + cos(\theta) \cdot H, y_{knee}, z_{knee} + sin(\theta) \cdot H) \quad (12)$$

where $x_{knee}, y_{knee}$ and $z_{knee}$ represents coordinate of the corresponding knee, $\theta$ is an angle parameter, and $H$ is the distance to target point. The target position is set when feet clipping is about to occur, and the leg movement is calculated using inverse kinematics. If no clipping is detected, the target position is ignored, and the model follows its original animation.

*3.3.2* ***Depth constraint***. In addition to the feet constraint, we introduce a depth constraint to optimize body jerking caused by incorrect camera detection. Similar to the approach used in the feet constraint section, we apply a position constraint to the body's root joint, limiting the range of motion of the root. The range of motion for the body root should be determined manually based on the specific characteristics of the animation and the desired level of constraint. In our implementation, we define the range as a sphere with a diameter equal to twice the body thickness. We also allow the constraint to be violated if necessary by setting it to a soft constraint, which helps to create more natural and smooth movements that resemble real-life motion.

By combining the depth constraint with the previously introduced feet physics optimization, we achieve sufficient movement for estimating ground reaction force (GRF). The depth constraint ensures the body moves within a plausible range, while the feet constraint maintains proper contact with the ground. However, it's worth mentioning that due to the nature of these constraints, they do not completely prevent the feet from floating. In cases where floating feet are observed after the 3D pose estimation process, although adjusting the plane coordinates may serve as a potential solution, this approach should be considered a secondary measure rather than the primary method for addressing the issue.

## 3.4 Ground Reaction Force

In this section, we present our approach to estimating ground reaction force (GRF) after performing the necessary optimizations. As proposed by Thompson [Thompson 2002], the ground reaction force is equal in magnitude and opposite in direction to the force that the body exerts on the supporting surface through the foot. Figure 5 illustrates the proposed diagram for ground reaction force, where $F = mg$ represents the force vector of gravity acting on the

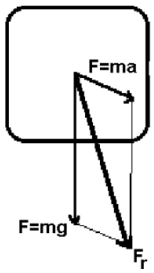



**Figure 5: Ground Reaction Force $F_r$**

object, and $F = ma$ represents the force vector of instantaneous inertial force acting on the object. The resultant force vector ($F_r$) is the sum of the gravitational and inertial forces, as formulated in the following equation:

$$GRF(F_r) = mg + ma \quad (13)$$

where $m$ is mass, $g$ is gravity, and $a$ is acceleration.

It is important to note that friction force can be considered as the third component of ground reaction force. However, measuring friction force is very challenging due to the complexity of the foot-ground interface and the limitations of current force platform technology. As a result, friction force is often ignored in biomechanical studies, as reported by [Damavandi et al. 2012][Morio et al. 2009]

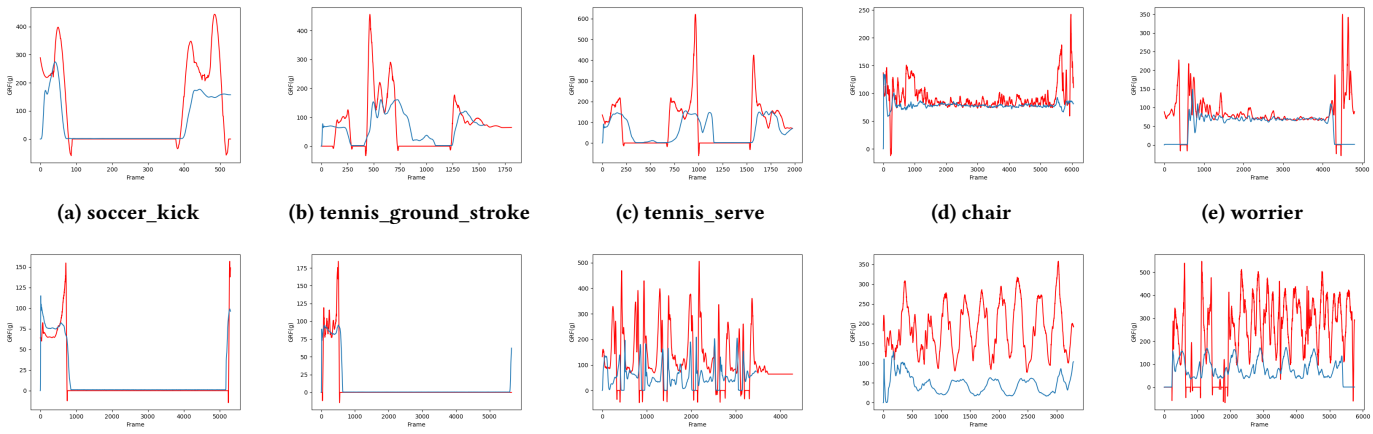
To adapt the ground reaction force calculation to our animation, we compute the force in every frame where the laser and collision detector detects contact. The mass of the object is mainly based on assumption, but real mass is used if it is provided. Calculating the acceleration is a critical component of the GRF estimation. Initially, we considered using only the root joint to calculate the change in position per frame. However, this approach was found to be sensitive only to dynamic movements.

To capture a more comprehensive representation of the leg movement and determine the acceleration accurately, we incorporate the entire lower body and root joint in the calculation. The change in position for acceleration is computed using the following equation:

$$a = N(\frac{1}{F}\sum_{F=1}^{F}(\frac{root + knee + ankle + toe}{4})) \quad (14)$$

where $F$ is the number of frames and $N$ is the normal distribution. By setting $F = 5$, we calculate the mean of the average lower body's joints change in position and normalize the acceleration for every 5 frames. This approach ensures a smooth transition in velocity and provides a more robust estimation of the acceleration.

By combining the optimized 3D poses, contact detection, and the adapted ground reaction force calculation, we can estimate the GRF for each frame of the animation. The proposed method provides a practical solution for estimating GRF in animations, enabling further analysis and applications in biomechanics and related fields.



(a) soccer_kick (b) tennis_ground_stroke (c) tennis_serve (d) chair (e) worrier

(f) tree (g) tree_arms (h) ballet_small_jump (i) dog (j) side_stretch

**Figure 6: Comparison against ground truth from GroundLinkNet dataset [Han et al. 2023] for 10 randomly chosen motions. Horizontal axis: Frame; vertical axis: Ground Reaction Force(g).**

## 4 RESULTS

In this section, we present the results of our proposed approach for estimating ground reaction forces (GRFs) from monocular video. We evaluate the performance of our method both quantita ively and qualitatively, comparing the estimated GRFs with ground truth data from the GroundLink dataset[Han et al. 2023].

It's important to note that due to time constraints and limited computational resources, we were unable to conduct a full training of our proposed 3D pose estimation network with the additional loss terms. Instead, we utilized the pretrained weights of the HybrIK-X model [Li et al. 2023] and applied our physics-based optimization and GRF estimation pipeline to the recovered 3D poses. While this limitation prevents us from fully assessing the potential improvements brought by our proposed modifications to the 3D pose estimation stage, we believe that our results still provide valuable insights into the effectiveness of our overall approach.

### 4.1 Qualitative Comparison

Overall, from the qualitative comparison figures, it is clearly evident that our calculated GRFs exhibit similar trends to the ground truth data. The estimated force profiles capture the key characteristics and timing of the GRF patterns, demonstrating the effectiveness of our physics-based optimization and GRF estimation pipeline.

Figure 6 (a) and (b) illustrates the estimated and ground truth GRF curves for the soccer kick motion, tennis ground stroke motion respectfully. The estimated GRF profile closely follows the overall shape and timing of the ground truth data, capturing the key peaks and valleys corresponding to the different phrases of both motions. However, the Magnitude of the estimated forces is slightly higher than the ground truth particularly during the impact phase of the motions.

For steady poses, our method demonstrates a high level of accuracy. Figures 6 (d) and (e) showcase the results of our method applied to continuous chair and warrior poses, respectively. In these cases, our approach successfully captures the stationary nature of the poses, accurately estimating the consistent force profiles during the sustained postures. Moreover, our method effectively captures the force transitions during the entering and exiting phases of the poses.

It is also worth noting that some failure cases can be observed in the results. These failures primarily occur in slow weight transfer poses, such as the dog pose and side stretcha, where the movement of the lower body joints is very gradual. In these cases, the slow joint velocities can lead to inaccuracies in the calculation of the velocity term in the GRF equation, resulting in miscalculations of the forces. Figure 6 (i) and (j) showcases the GRF comparison for the side stretch motions and dog pose motion. The estimated forces are higher than ground truth in average for both motions, since the force distributed between other points of contact are not captured.

### 4.2 Quantitative Comparison

To evaluate the accuracy of our estimated GRFs, we compare them with the ground truth GRF data provided in the GroundLink dataset. Table 1 reports the mean squared error (MSE) of the calculated ground reaction forces from our proposed approach for both the left and right feet, considering 10 randomly selected motions.
It is important to acknowledge that, to the best of our knowledge, this is the first paper to utilize a Unity implementation for calculating GRFs. As a result, there is limited availability of directly comparable data from prior works. This highlights the novelty of our approach and the potential for further exploration in this area.

The results in Table 1 provide insights into the performance of our method across different motion types. The MSE values range from 0.003 to 0.150, indicating varying levels of accuracy in the estimated GRFs. Lower MSE values, such as those observed for the tree arms and tree motions, suggest a closer match between the estimated and ground truth forces. On the other hand, higher MSE values, like those seen for the side stretch and dog motions, indicate larger discrepancies between the estimated and true forces.

| Motions | *Left Leg* | *Right Leg* |
|---|---|---|
| soccer_kick | 0.051 | 0.043 |
| dog | 0.106 | 0.112 |
| chair | 0.013 | 0.028 |
| side_stretch | 0.150 | 0.148 |
| tree_arms | 0.003 | 0.006 |
| tree | 0.004 | 0.011 |
| worrier | 0.017 | 0.009 |
| tennis_serve | 0.023 | 0.005 |
| ballet_small_jump | 0.078 | 0.069 |
| tennis_ground_stroke | 0.029 | 0.016 |

**Table 1: Mean Square Error with GroundLink ground truth dataset**

These variations in MSE can be attributed to the unique characteristics of each motion. Motions with more dynamic and rapid movements, such as the soccer kick and tennis serve, tend to have higher MSE values compared to more static or slow-moving poses like the tree arms and tree. This observation aligns with the previously discussed challenges in estimating GRFs for slow weight transfer poses.

Furthermore, the MSE values for the left and right feet within each motion are generally comparable, suggesting a consistent performance of our method in estimating GRFs for both limbs. However, there are a few cases where the MSE differs slightly between the left and right feet, which could be due to asymmetries in the motion or variations in the quality of the 3D pose estimates.

In summary, our results demonstrate the potential of our proposed approach for estimating GRFs from monocular video. The quantitative evaluation reveals promising accuracy across various motion types, with room for improvement in handling slow weight transfer poses and refining the force magnitudes. The qualitative comparison showcases the ability of our method to capture the key characteristics and trends of the GRF profiles, providing valuable insights into the dynamics of human motion. Despite the limitations imposed by the use of pretrained 3D pose estimates, our results highlight the effectiveness of our physics-based optimization and GRF estimation pipeline.

## 5 PROJECT MANAGEMENT

[gaunt chart] []

## 6 CONCLUSION

In this paper, we proposed a pipeline for Ground Reaction Force Estimation from Monocular Video. Our approach begins by optimizing feet contact points using a newly proposed feet adjustment function, which is then combined with physics-based position constraints to further refine the error in the contact points of the feet. We evaluated our method using the Groundlink dataset [Han et al. 2023] by calculating the mean squared error differences between our estimated GRFs and the ground truth values. The results indicate that our pipeline achieves fairly accurate GRF estimation, with a clear trend of force shifting captured in the estimated profiles. Although we were unable to train our model with the new loss function due to the extremely high computational requirements for GPU resources, we believe that our proposed pipeline represents a novel innovation that can significantly lower the difficulty of future GRF estimation tasks.

### 6.1 Limitation and Future Work

One of the main limitations we faced during this research was the lack of computational resources for training the neural network. As the requirements for achieving more accurate results continue to grow exponentially, this becomes a significant physical limitation that the entire field of computer vision needs to address. Additionally, our proposed method exhibits relative weakness in handling movements where the feet are constantly floating. While applying constraints is necessary for accurate pose estimation, setting appropriate constraints to allow for feet floating can be an ill-posed problem.

For future work, we suggest proceeding with our proposed function given sufficient physical resources. With further research and development, it is theoretically possible to perform contact point estimation entirely through deep learning training, eliminating the need for manual adjustments and constraints. By achieving this, we will be able to estimate GRFs from general movements captured in any monocular video, making the majority of videos sufficient to serve as ground truth. This advancement would prevent the need for expensive setups and specialized equipment for capturing GRFs, opening up new possibilities for biomechanical analysis, sports performance evaluation, and rehabilitation monitoring. However, continued efforts are needed to improve the robustness, generalization, and computational efficiency of GRF estimation algorithms to realize this vision.

## ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my supervisor, Hyung Jin Chang, for his invaluable guidance and support throughout this project. His expertise and insights were instrumental in shaping the research idea and providing the necessary resources to bring it to fruition. I am truly grateful for his mentorship and the knowledge he imparted, which have been crucial to the success of this work.

I also extend my heartfelt thanks to Esha Dasgupta, a PhD student at the University of Birmingham and a member of Hyung's lab. Esha's unwavering support and assistance have been indispensable to me throughout this project. From offering mental support and sharing her knowledge to suggesting approaches and helping with their implementation, Esha has been a constant source of encouragement and guidance. Her contributions have been essential to overcoming the challenges encountered during this research, and I cannot imagine completing this project without her dedication and support.

I am incredibly fortunate to have had the opportunity to work with both Hyung and Esha, and I am sincerely grateful for their mentorship, expertise, and unwavering commitment to this project.

## REFERENCES

Andrea Ancillao, Salvatore Tedesco, John Barton, and Brendan O'Flynn. 2018. Indirect Measurement of Ground Reaction Forces and Moments by Means of Wearable

Inertial Sensors: A Systematic Review. *Sensors* 18, 8 (2018). https://doi.org/10.3390/s18082564

Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. 2020. Beyond Static Features for Temporally Consistent 3D Human Pose and Shape from a Video. *CoRR* abs/2011.08627 (2020). arXiv:2011.08627 https://arxiv.org/abs/2011.08627

M. Damavandi, P. C. Dixon, and D. J. Pearsall. 2012. Ground reaction force adaptations during cross-slope walking and running. *Human Movement Science* 31, 1 (2012), 182–189. https://doi.org/10.1016/j.humov.2011.06.004

Kistler Group. 2020. Force plates. https://www.kistler.com/en/applications/sensor-technology/biomechanics-and-force-plate/. Accessed: YYYY-MM-DD.

Xingjian Han, Ben Senderling, Stanley To, Deepak Kumar, Emily Whiting, and Jun Saito. 2023. GroundLink: A Dataset Unifying Human Body Movement and Ground Reaction Dynamics. In *SIGGRAPH Asia 2023 Conference Papers (SA '23)*. ACM. https://doi.org/10.1145/3610548.3618247

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2017. End-to-end Recovery of Human Shape and Pose. *CoRR* abs/1712.06584 (2017). arXiv:1712.06584 http://arxiv.org/abs/1712.06584

Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. 2019. VIBE: Video Inference for Human Body Pose and Shape Estimation. *CoRR* abs/1912.05656 (2019). arXiv:1912.05656 http://arxiv.org/abs/1912.05656

Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. 2019. Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop. In *ICCV*.

Marek Kulbacki, Jakub Segen, and Jerzy Paweł Nowacki. 2021. 4GAIT: Synchronized MoCap, Video, GRF and EMG Datasets: Acquisition, Management and Applications. *arXiv preprint arXiv:2112.03553* (2021).

Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu. 2023. HybrIK-X: Hybrid Analytical-Neural Inverse Kinematics for Whole-body Mesh Recovery. arXiv:2304.05690 [cs.CV]

Jiefeng Li, Chao Xu, Zhicun Chen, Bian, Lixin Yang, and Cewu Lu. 2020. HybrIK: A Hybrid Analytical-Neural Inverse Kinematics Solution for 3D Human Pose and Shape Estimation. *CoRR* abs/2011.14672 (2020). arXiv:2011.14672 https://arxiv.org/abs/2011.14672

Kevin Lin, Lijuan Wang, and Zicheng Liu. 2020. End-to-End Human Pose and Mesh Reconstruction with Transformers. *CoRR* abs/2012.09760 (2020). arXiv:2012.09760 https://arxiv.org/abs/2012.09760

Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. 2018. PlaneRCNN: 3D Plane Detection and Reconstruction from a Single Image. *CoRR* abs/1812.04072 (2018). arXiv:1812.04072 http://arxiv.org/abs/1812.04072

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

C. Morio, M. J. Lake, N. Gueguen, G. Rao, and L. Baly. 2009. The influence of footwear on foot motion during walking and running. *Journal of Biomechanics* 42, 13 (2009), 2081–2088. https://doi.org/10.1016/j.jbiomech.2009.06.015

Lucas Mourot, Ludovic Hoyet, François Le Clerc, and Pierre Hellier. 2022. UnderPressure: Deep Learning for Foot Contact Detection, Ground Reaction Force Estimation and Footskate Cleanup. arXiv:2208.04598 [cs.GR]

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Davis Rempe, Leonidas J. Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. 2020. Contact and Human Dynamics from Monocular Video. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Erfan Shahabpoor and Aleksandar Pavic. 2017. Measurement of Walking Ground Reactions in Real-Life Environments: A Systematic Review of Techniques and Technologies. *Sensors* 17, 9 (2017). https://doi.org/10.3390/s17092085

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5693–5703.

D. Thompson. 2002. Ground reaction force. University of Oklahoma Health Sciences Center. https://ouhsc.edu/bserdac/dthompso/web/gait/kinetics/GRFBKGND.HTM Accessed: 2024-04-03.

Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. 2023. 3D Human Pose Estimation via Intuitive Physics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4713–4725.

Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. 2021. Physics-based Human Motion Estimation and Synthesis from Videos. *CoRR* abs/2109.09913 (2021). arXiv:2109.09913 https://arxiv.org/abs/2109.09913

Guanyu Yang, Wen-Hao Hsu, Kevin Chou, Jiajun Hu, Jiajun Wu, Daniel Yamins, and Taku Komura. 2021. From Image to Stability: Learning Dynamics from Human Pose. *arXiv preprint arXiv:2109.14076* (2021).

Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. 2022. Physical Inertial Poser (PIP): Physics-aware Real-time Human Motion Tracking from Sparse Inertial Sensors. arXiv:2203.08528 [cs.GR]

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. 2023. PhysDiff: Physics-Guided Human Motion Diffusion Model. arXiv:2212.02500 [cs.CV]

Yeqing Zhu, Di Xia, and Heng Zhang. 2023. Using Wearable Sensors to Estimate Vertical Ground Reaction Force Based on a Transformer. *Applied Sciences* 13, 4 (2023). https://doi.org/10.3390/app13042136