2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Monocular 3D Reconstruction of Cheetahs in the Wild*

Zico da Silva[1], Zuhayr Parkar[1], Naoya Muramatsu[1], Fred Nicolls[1], Amir Patel[1†]

***Abstract*— This paper introduces a framework for monocular 3D reconstruction of cheetah movements, leveraging a combination of data-driven and physics-based modeling as well as trajectory optimization. Unlike traditional methods that rely solely on kinematics, our approach integrates dynamic motion principles, enhancing the plausibility and generalization of motion estimates. Validated on the cheetah running dataset, AcinoSet, we achieve mean per-joint position errors of 78.8 mm and 72.5 mm, showcasing significant advancements over the existing model used in AcinoSet. By addressing the challenge of absent ground truth data, this work not only advances animal motion capture techniques but also informs the development of bio-inspired robotic systems, offering a robust solution for accurately capturing complex animal locomotion in natural settings.**

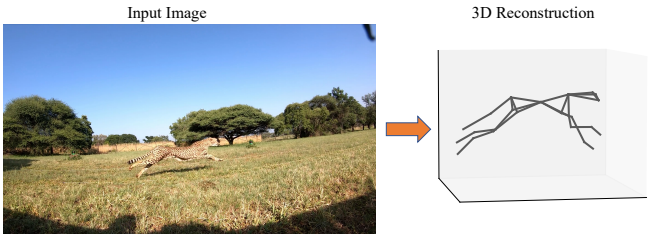


Fig. 1: Cheetah in flight phase of locomotion taken from AcinoSet [2] and its corresponding 3D reconstruction using a monocular image.

I. Introduction

Studying animal movements in their natural habitats is essential for understanding how they navigate complex environments rapidly. The cheetah (*Acinonyx jubatus*), known for its unparalleled speed and agility, serves as a uniquely challenging subject for examining quadruped dynamics. Through 3D motion capture of cheetahs in the wild, researchers can unravel the biomechanics of these animals, offering insights critical for the development of legged robots that navigate complex terrains more effectively than their wheeled counterparts [1].

This research not only advances robotic mobility but also contributes to the conservation of wild cheetahs. Traditional methods for in-field 3D motion capture, especially for high-speed wildlife like cheetahs, are often too invasive. Recent efforts have aimed to overcome this by employing non-invasive, markerless 3D motion capture using multi-view low-cost camera systems to create the first cheetah running dataset, AcinoSet [2]. However, the complexity of multi-camera setups limits the flexibility of motion capture data that can be collected.

A solution to this limitation lies in adopting a monocular camera system. Such a system simplifies the capture process significantly—reducing the need for costly equipment and intricate calibration, and enabling the extension of the capture area to track moving subjects. Thus, a monocular system facilitates the collection of a more diverse range of data and allows for the utilization of existing wildlife footage. An example of a 3D monocular reconstruction is shown in Fig. 1.

To address the challenges of 3D monocular reconstruction, we investigate two 2D-to-3D methods using both data-driven and physics-based modelling that builds on the full trajectory estimation (FTE) method developed for AcinoSet. These methods leverage a prior model of the cheetah's pose and motion to enhance the accuracy of 3D reconstruction. Our work includes improvements to AcinoSet using a previously developed multi-camera system. This facilitates the training of a data-driven model through linear regression and a Gaussian mixture model. Furthermore, this aids in the evaluation of the the monocular 3D reconstruction results. Lastly, we adopt our physics-based model from a similar FTE method developed in our previous work on dynamic estimation [3].

Our findings demonstrate that both models significantly improve 3D reconstruction accuracy on a subset of AcinoSet, with mean per-joint position errors (MPJPE) of 78.8 mm and 72.5 mm, respectively. These improvements reflect a 50 % reduction in error for straight-line runs and a 20 % reduction for more complex maneuvers. Although both models exhibit comparable performance, the physics-based approach demonstrates better generalization and produces more physically realistic estimates, while the data-driven model achieves faster solution times.

II. Literature Review

A. Data-driven Models

Data-driven approaches in monocular 3D motion estimation prioritize learning from training sets, typically derived from motion capture systems. Historically, statistical modeling, particularly within a Bayesian framework, was prevalent in the late '90s for human motion analysis [4]–[6]. These methods frequently employed Principal Component Analysis (PCA) for reducing the dimensionality of motion data, creating a more manageable subspace without sacrificing critical information [7]. Studies like [4] and [6] utilized Gaussian probability distributions over PCA-obtained subspaces for

This work was supported by The National Research Foundation of South Africa (NRF), Grant no. 137762 and the Google Research Scholar Program.
†The corresponding author amir.patel@uct.ac.za.
[1] African Robotics Unit (ARU), University of Cape Town, South Africa

statistical modeling, while [5] applied a Gaussian mixture model (GMM) to encapsulate prior probability distributions.

The inherent symmetry and repetitiveness in human and animal motion render PCA and Markov processes, such as hidden Markov models (HMM), effective for constraining the solution space and constructing motion models [8]. This synergy has been notably applied in human motion synthesis, particularly within the computer graphics domain [8]–[10].

The advent of large, diverse datasets has catalyzed deep learning's application in 3D pose estimation, facilitating direct 3D pose estimation from 2D data through methods like pose "lifting" [11] and CNN-based regression processes [12]. These advancements have extended to laboratory animal studies, exemplified by the LiftPose3D toolbox [13], despite its limitations in capturing temporal dynamics and the added complexity of natural environments for generating reliable ground truth data. This backdrop underscores our data-driven approach's foundation, which builds upon principles established before the surge of deep learning, aiming to address the unique challenges posed by natural environment motion capture.

B. Physics-based Models

Data-driven approaches often utilize kinematic models that overlook kinetic motion's foundational principles, potentially leading to biomechanically inaccurate and overly generalized motion estimates. Physics-based models address these limitations by integrating kinetic principles—such as internal forces and reaction to external forces—providing a more robust framework for dynamic motion analysis. This integration not only enhances motion plausibility but also facilitates the detailed study of biomechanics through the evaluation of internal joint torques and external ground reaction forces, an approach increasingly adopted for human torque analysis in optimal control problems [14], [15].

Trajectory optimization has emerged as a preferred method for refining monocular 3D pose estimation, offering significant advancements by incorporating physics-based modeling [16]–[20]. Notably, this approach was utilized to analyze dynamic human-object interactions from internet videos, providing insights into complex motion patterns [17].

However, trajectory optimization methods often face challenges in real-time execution due to the complexity of solving large-scale non-linear optimization problems. Alternative physics-based methods that do not require full trajectory consideration offer real-time capabilities but rely heavily on accurate initial kinematic reference motion, achieved through PD controllers or reinforcement learning policies [21]–[24]. Additionally, the majority of these approaches necessitate a prior knowledge of contact timing, often determined via neural network architectures that estimate contact states based on 2D poses [17], [21]. More recently, differentiable physics simulators have been used for monocular human pose estimation that do not need known contact timing [25].

In our work, we employ the trajectory optimization framework, echoing the non-real-time methodology used in the multi-camera system study of [2], with the aim to refine its application without the requirement for real-time processing. Given the constraints of relying on precise initial kinematic estimates for real-time simulation, we focus on the more comprehensive trajectory optimization approach, setting aside the discussion on real-time methods due to feasibility concerns.

III. Dataset

A. AcinoSet

AcinoSet [2] is a cheetah running dataset that is used for the evaluation of all methods in this work. The dataset contains 90 running videos with six different views. There are 7588 human-annotated frames and the average video length is approximately two seconds.

Out of a total of 93 trials in AcinoSet, only 27 were deemed reliable enough to be used as "ground truth". Trials were chosen based on:

- visual plausibility of the resultant 3D reconstruction,
- the accuracy of the estimated multi-view camera extrinsic parameters, and
- whether the multi-view camera system was synchronized correctly for a particular trial.

Left with 27 trials, the data was split into a training set of size 15 and a test set of size 12, corresponding to 1629 and 1340 individual poses respectively. This was further reduced to 1539 and 1268 individual poses by removing the first and last three poses for each trial, since it was evident that the start and end poses of a reconstruction suffered from edge effects.

The training set was used to estimate the prior pose and motion model described in Section V-A. The test set used to evaluate the performance of each monocular method is shown in Table I. It is a subset taken from AcinoSet for two different cheetahs, consisting of 10 trials: 5 steady-state runs and 5 maneuvers. A steady-state run is defined as a straight continuous run at a constant speed. A maneuver is defined as anything that cannot be classified as a steady-state run, i.e. runs that involve turns or deceleration/acceleration events. Note that each trial contains one stride of the cheetah.

TABLE I: The test dataset selected from AcinoSet. The length is presented as the number of frames in the trajectory divided by the video frame rate.

Test	Trial	Length (s)
Manoeuvres:		
T1	2017_12_09/bottom/jules/flick2	30/90
T2	2019_03_09/jules/flick1	34/120
T3	2019_03_03/phantom/run	42/120
T4	2017_09_02/top/phantom/run1_2	45/90
T5	2017_08_29/top/jules/run1_2	34/90
Runs:		
T1	2017_08_29/top/phantom/run1_1	44/90
T2	2017_08_29/top/jules/run1_1	30/90
T3	2017_09_02/top/jules/run1	30/90
T4	2019_03_07/phantom/run	57/120
T5	2017_09_02/bottom/jules/run2	33/90

B. Kinetic Dataset

The kinetic dataset was used to develop and evaluate the use of a physics-based model of the cheetah [3]. It was acquired from the Royal Veterinary College [26]. The original work done for [26] gathered trials of the cheetahs chasing a mechanical lure across eight force plates and in front of four cameras.

From the original dataset, a subset of 5 trials of two different subjects[1] was selected. Both the video and the force plate data were resampled to a 200 Hz sample rate to be used in our experiment.

IV. Multi-body Dynamics

The state vector of a cheetah pose is defined as $\mathbf{q} \in \mathbb{R}^{3+3L}$, which contain absolute coordinates for the angles, to fully describe the configuration of the system. Note that L is the number of links in the system and $\mathbf{q}$ denotes all positional information relative from the center of mass (COM) about the system.

In this work, two different dynamic functions are used: kinematic and kinetic. The kinematic model does not consider the forces that produced the motion, whereas the kinetic model includes internal and external forces acting within and on the system. The kinematic model assumed a constant acceleration between time steps, allowing the dynamics to be simply captured as

$$\ddot{\mathbf{q}}(k+\Delta k) = \ddot{\mathbf{q}}(k) + \mathbf{w}(k), \tag{1}$$

where $\ddot{\mathbf{q}}(k)$ denotes the accelerations of the system at time k and $\mathbf{w}(k)$ is an input disturbance that allows for deviations from the assumed constant model. This is an important quantity for estimating motion, because it is not certain that our motion model perfectly captures the dynamics of the subject—instead, deviations from this model are allowed, but minimized as much as possible.

The kinetic model follows the more general form for the equations of motion of a rigid multi-body system (below the time parameter k has been omitted for clarity, but it is assumed that this equation is time dependent), which is often expressed as

$$\mathbf{M}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q},\dot{\mathbf{q}})\dot{\mathbf{q}} = \mathbf{G}(\mathbf{q}) + \mathbf{B}\mathbf{u} + \mathbf{J}_L^T(\mathbf{q})\boldsymbol{\lambda} + \mathbf{J}_A^T(\mathbf{q})\mathbf{a} + \mathbf{w}, \tag{2}$$

where $\mathbf{M}(\mathbf{q})$ represents the inertia matrix, $\mathbf{C}(\mathbf{q},\dot{\mathbf{q}})$ captures Coriolis and centrifugal terms, $\mathbf{G}(\mathbf{q})$ is the gravity vector, $\mathbf{B}$ maps inputs $\mathbf{u}$ to generalized forces, $\mathbf{J}_L^T(\mathbf{q})$ is the contact Jacobian and $\boldsymbol{\lambda}$ the corresponding contact forces, and $\mathbf{J}_A^T(\mathbf{q})$ is the angle constraint Jacobian that maps the constraint torques $\mathbf{a}$ into the equation of motion.

In this work $\mathbf{Bu}$ represents the joint torques produced by the cheetah during locomotion, henceforth referred to as $\boldsymbol{\tau}$. For both dynamic equations (Equations 1-2), the position and velocity (i.e. the state of the system) are easily determined from the accelerations by using Euler's method, as shown:

$$\dot{\mathbf{q}}(k+\Delta k) = \dot{\mathbf{q}}(k) + \Delta k\ddot{\mathbf{q}}(k+\Delta k), \tag{3}$$

$$\mathbf{q}(k+\Delta k) = \mathbf{q}(k) + \Delta k\dot{\mathbf{q}}(k+\Delta k). \tag{4}$$

V. Data-driven Approach

AcinoSet was used to facilitate data-driven techniques to obtain a strong prior on the pose and motion of the cheetah. The pose prior was modeled using a GMM and the motion prior was modeled using linear regression (LR).

A. Pre-process

Both of our models require that the dataset be manipulated to isolate the data of interest. The pose prior is estimated using a GMM to model the distribution of the training dataset. For this purpose the absolute position and orientation of the cheetah is not considered, as it provides no useful information about its pose. Therefore, the pose vector is truncated to exclude these parameters (from size 28 to 22).

The motion prior is assumed to be a linear model that predicts the evolution of the pose state. This predictor function is estimated using LR with a window size, T_w, that encapsulates the number of past states used to predict the next. For example, a regression function $\mathbf{q}_k^r = g(\mathbf{q}_{k-1}^r, \mathbf{q}_{k-2}^r)$ has a window size $T_w = 2$ to predict the current pose at time k from the previous two poses. The training set without manipulation is represented as $\mathbf{X} \in \mathbb{R}^{1539\times 28}$. In order to obtain an estimate of the regression function above, the dataset is transformed into the form $\mathbf{X} \in \mathbb{R}^{1539\times 28T_w}$. This provides the necessary structure to perform time series forecasting using LR.

B. Full Trajectory Estimation

The goal is to "score" the trajectory with both data-driven models, i.e. determine the likelihood of an output trajectory. In doing so, good solutions are expected to score high for both models, resulting in outputs that are similar to those found in the training set. The idea is that the log of the likelihood function can be used to measure the "goodness of fit" for a particular model. This log likelihood function then serves as an error metric that is minimized in the cost function

$$g(\mathbf{q}) = e_{meas} + e_{model} + e_{pose} + e_{motion}, \tag{5}$$

where e_{meas} and e_{model} terms have been established in [27]. The e_{pose} term is modeled using the GMM and e_{motion} is modeled using a regression function. Each term has been normalized independently and therefore has equal importance during minimization.

For the GMM, the negative log-likelihood allows the determination of the pose likelihood. For example, the higher the value the more unlikely the pose, given the model. The negative log-likelihood of the GMM is defined as

$$e_{pose} = \sum_{k=1}^{N} -\log(\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}|\mathbf{q}_k)), \tag{6}$$

[1] The subjects were *Cheetah 1* and *Cheetah 2* referred to in Table 1 of [26].

where $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\alpha}$ are learned parameters of the GMM and $\mathcal{L}$ denotes the likelihood function.

The linear regression function is of the form, $\mathbf{y} = \boldsymbol{\beta}^T \mathbf{q}^r + \boldsymbol{\epsilon}$. This is of a similar form to the dynamic function $f(\cdot)$, where the current state, along with some added noise, is used to predict the next state. If the noise is assumed (in this case $\boldsymbol{\epsilon}$) to be normally distributed (a valid assumption when using LR to obtain the model), the negative log-likelihood can be determined by the sum of squared errors:

$$e_{motion} = \sum_{i=1}^{N} \sum_{j=1}^{p_r} \left(\frac{\boldsymbol{\epsilon}_{i,j}}{\sigma_{motion,j}} \right)^2, \tag{7}$$

where $\sigma_{motion,j}$ is the standard deviation of the prediction error on the training set, and p_r is the size of the pose vector (i.e. 28). In addition, it is assumed that each component j is independent, i.e. the covariance estimate from the trajectory is diagonal.

Similar to the dynamic function, an equality constraint is established to capture $\boldsymbol{\epsilon}$:

$$\mathbf{y}_k - \boldsymbol{\beta}^T \mathbf{q}^r_{k-1:k-T_w} - \boldsymbol{\epsilon}_k = \mathbf{0}, \tag{8}$$

defined at each finite time k.

VI. Physics-based Approach

The main goal for the kinetic dataset was to determine the plausibility of adopting a physics-based model proposed in [3], where the contact timing was assumed to be given for simplicity. However, here the contact state and timing need to be determined automatically in order to develop a method that is comparable to the data-driven approach. Hence contact detection is required. This operation is performed prior to the physics-based FTE, which allows for the original formulation of the physics-based FTE to be used thereafter. Here, the original physics-based FTE formulation is constrained to a single camera, i.e. $c = 1$, and the cost function includes the GMM pose prior. The resultant cost function for the monocular physics-based FTE is

$$g(\mathbf{q}, \boldsymbol{\tau}) = \alpha_1 e_{meas} + \alpha_2 e_{model} + \alpha_3 e_{pose} + \alpha_4 e_{smooth}, \tag{9}$$

where e_{smooth} is established in [3], and $\alpha_1 = 1$, $\alpha_2 = 10\,000$, $\alpha_3 = 1$, and $\alpha_4 = 1$. Each hyper-parameter was chosen through experimentation. The pose prior reduces the likelihood of invalid pose estimates, as it did in the data driven model.

The kinematic FTE solution was determined so that contact detection is performed on the resultant trajectory. A simple height heuristic (when the foot is below a certain threshold it is assumed to be in contact with the ground) to determine the contact points cannot be relied upon for the following reasons:

- The constant acceleration model has a smoothing effect on the resultant trajectory. This in turn makes it difficult to know the duration of the contact, if its discontinuous characteristics are removed by the smoothing.
- The rigid body model of the cheetah potentially exacerbates the previous issue, by preventing the common oscillatory behavior of shoulder joints during contact. This results in additional uncertainty in the determination of the touchdown and takeoff contact events.

Thus, a more nuanced contact detection algorithm is required. From [26], a contact detection process was developed to take the cheetah's speed into account to determine the duration of the stance using a linear model. As such, if a touchdown event is detected and the speed of the cheetah is known, a prediction of the takeoff event is possible. Hence the time and duration of the contact has been determined.

The timing of when a touchdown event occurs is very challenging to ascertain, especially given the smooth kinematic FTE solution. Consequently, in implementation, a height heuristic (when the cheetah's paw is within 5 cm of the ground) together with a zero-crossing velocity event is used to determine the rough location of the midpoint of a stance. Then the duration of the stance was used to complete the contact detection process.

A. Contact Detection

Table II provides the results of the contact detection method. For the trials used in both the kinetic and AcinoSet datasets, the contact points were labeled using the start and end times of each contact event, forming a ground truth to compare against. Note that with the kinetic dataset the paws are clearly visible at touchdown and takeoff contact events. This is not the case for AcinoSet, and therefore uncertainty is attached to the quoted result in Table II. The assumption is that a contact point is reliably labeled to within ± 1 frames. For AcinoSet, this corresponds to a total of 40 uncertain contact points (10 examples of 4 contact events)—or in other words 13.6 % (40 out of a total of 294 labeled points).

As expected, evaluation on the kinetic dataset yields the highest success rate due to the clear visibility of the paws. Also, the linear model used to relate the cheetah's speed with the stance duration was originally determined using this dataset in [26]. Evaluation on AcinoSet was done for trajectories obtained via the kinematic and data-driven FTE (referred to as 'Monocular AcinoSet' in Table II) methods. Surprisingly, the contact detection using monocular reconstructions obtain a better success rate than when using multiple views. Nevertheless, the confidence bands do overlap, meaning that a comparison has little significance. However, with certainty, the contact detection performance on AcinoSet is definitely in excess of 60 %, which appeared to be adequate for our purposes.

TABLE II: Contact detection results on both datasets.

Dataset	# Labeled	# Estimated	Success Rate (%)
Kinetic	201	216	90.3
Multi-view AcinoSet	294	300	75.3 ± 13.6
Monocular AcinoSet	294	292	83.0 ± 13.6

VII. Evaluations

Two different approaches have been established to perform monocular 3D reconstructions of the cheetah in the wild. Here, a quantitative analysis of the performance of

each method on the test set is presented, together with a comparison between the kinematic (referred to as "Default" in the results), data-driven, and physics-based FTE methods.

A. Comparative Study

The results for the steady-state runs are shown in Table III and for the maneuvers in Table IV. We evaluated them with three scores: mean position error (MPE), mean per-joint position error (MPJPE) and COM velocity error (CVE). In Table III it is clear that both the data-driven and physics-based approaches outperform the default case in every category. Both approaches reduce the error by approximately 50 % for every category, which in itself is a significant improvement. The data-driven approach has the lowest value for MPE, while the physics-based approach has the lowest MPJPE and CVE scores. The difference in average MPE is only 6.9 mm and 1.1 mm for the MPJPE.

In Table IV, the same can be said about both approaches outperforming the default case in every category, albeit not as significantly. Here, the error is reduced by approximately 20 %. The physics-based approach scores the best for the MPJPE and CVE, while the data-driven approach has the lowest MPE. As with Table III, it is doubtful whether one approach clearly outperforms the other, even though there is a 60.4 mm difference in MPE. There is an outlier value for the physics-based T3 in Table IV, which explains this disparity.

In conclusion, the relative tracking of joints is similar between the two approaches; however, the data-driven approach has an advantage in tracking the absolute translations, while the physics-based approach provides better accuracy in tracking the dynamics of the COM. This suggests that motion tracking of the cheetah is better done with the physics-based approach, even though the data-driven approach stays closer to the cheetah in 3D space.

Overall, the performance on the steady-state run dataset yields a significantly lower error than the maneuver dataset. This is expected, as the assumption is that maneuvers are a more challenging estimation problem.

Lastly, it should be mentioned that there is a possibility of an inherent bias towards the data-driven and default approaches because the constant acceleration motion model is shared across these methods, including the ground truth data. Therefore, the physics-based approach might unfairly be penalized for its use of a different dynamic equation of motion.

The average optimization time for each method is shown in Table V[2]. The physics-based approach takes significantly longer to perform the estimation task. This is consistent with the literature on the limitations of a physics-based trajectory optimization approach for motion estimation [23].

B. Ablation Study

For both approaches, there were added terms in the cost function to regularize the solution. The contribution of each term is evaluated by systematically removing terms from the cost function. This analysis is summarized in Fig. 2. The 'pose' refers to the pose term e_{pose}, and the 'motion' refers to the motion term e_{motion} for the data-driven approach and the smooth term e_{smooth} for the physics-based approach.

It is clear from Fig. 2 that the inclusion of both pose and motion terms produce the best result for all metrics. Consequently, both terms combine to great effect. For the data-driven approach, the pose and motion terms appear to have a similar influence on the MPE and MPJPE. However, as anticipated, the motion term provides a greater impact on the CVE compared with the sole inclusion of the pose term. In the physics-based approach, the pose term justifies its inclusion by effectively reducing both the MPE and MPJPE, while the motion term provides little influence in this respect—as is expected. On the whole, Fig. 2 confirms the benefit of including both terms into the cost function for both approaches.

C. Qualitative Results

Visual evaluation of the results aids the interpretation of the quantitative results in the previous sections. The side and top views of three example trials are shown in Figures 3-5, with the red circles indicating considerable errors in pose. Note that the orange cheetah skeleton in the top view denotes the monocular estimate, and the black cheetah skeleton is the ground truth.

A visual example of the trial that produced one of the lowest MPE and MPJPE values in Table III is shown in Fig. 3. Evidently, the quantitative analysis is consistent with

TABLE III: Comparison of monocular 3D reconstruction on the steady-state run dataset. Bold values are representative of the 'best' result in that category.

		T1	T2	T3	T4	T5	Avg
MPE (mm)	Default	158.9	351.5	185.5	171.3	395.8	252.6
	Data-driven	**109.0**	**133.6**	**136.7**	135.7	**133.0**	**129.6**
	Physics-based	126.2	137.8	159.5	**97.7**	161.3	136.5
MPJPE (mm)	Default	88.4	134.4	123.1	95.5	176.7	123.6
	Data-driven	**49.7**	64.7	**64.5**	**55.9**	80.7	63.1
	Physics-based	57.6	**59.2**	64.9	64.4	**63.9**	**62.0**
CVE (m)	Default	0.74	1.16	1.49	0.54	2.96	1.38
	Data-driven	0.42	**0.35**	0.78	0.57	0.56	0.54
	Physics-based	**0.36**	0.55	**0.73**	**0.18**	**0.45**	**0.45**

TABLE IV: Comparison of monocular 3D reconstruction on the manoeuvres dataset. Bold values are representative of the 'best' result in that category.

		T1	T2	T3	T4	T5	Avg
MPE (mm)	Default	430.8	195.8	346.7	147.7	260.5	276.3
	Data-driven	246.8	**178.5**	**218.1**	165.3	**108.8**	**183.5**
	Physics-based	**109.0**	260.5	531.2	**122.8**	201.3	243.9
MPJPE (mm)	Default	173.6	117.1	**77.8**	91.5	106.4	113.3
	Data-driven	115.2	103.7	98.8	79.2	**75.8**	94.5
	Physics-based	**80.7**	**89.6**	85.8	**66.6**	92.4	**83.0**
CVE (m)	Default	1.49	1.01	2.62	1.01	1.80	1.59
	Data-driven	1.5	**0.55**	**0.71**	0.68	0.85	0.86
	Physics-based	**0.48**	0.77	0.74	**0.21**	**0.49**	**0.54**

TABLE V: Optimization time for each method.

Method	Average Time (s)
Default	20
Data-driven	26
Physics-based	726

[2]Experiments were run on an Intel Core i9-7980XE processor (18 cores, 36 threads, 2.60 GHz base clock speed) and 32 GB of RAM.

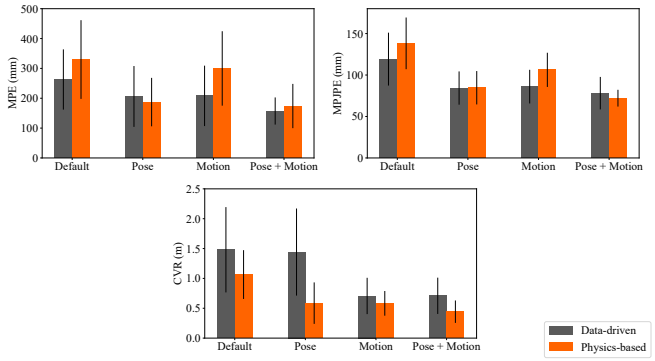


Fig. 2: Ablation study results that clearly show the contribution of each component added to the cost function. Error bars are added to capture variability.

the visual results. There is a single incorrect pose Fig. 3(a) and some unnatural body sliding Fig. 3(b) for the default method. Otherwise, both the data-driven and physics-based methods closely resemble the ground truth.

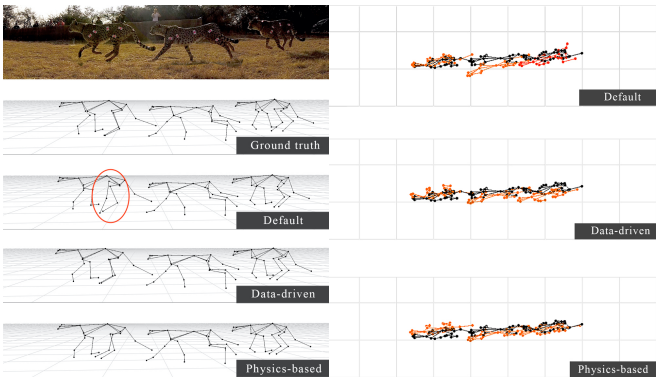
A visual example of the trial that produced the biggest difference in MPE between the physics-based and data-driven approaches, in favor of the physics-based approach (Table IV), is shown in Fig. 4. It is evident in Fig. 4(b) that both the default and data-driven approaches fail to produce a physically plausible trajectory, whereas the physics-based approach appears to track the motion of the cheetah with reasonable accuracy. This clear incorrect reconstruction is not as obvious in Fig. 4(a).

A visual example of the trial that produced the biggest difference in MPE between the physics-based and data-driven approaches, in favor of the data-driven approach (Table IV), is shown in Fig. 5. The physics-based approach produced a high MPE because it is clearly off the ground truth track for much of the trajectory. However, it roughly performs the same motion, albeit at a slightly different location in 3D space. This is in contrast to the other methods that might favour implausible 3D translations to lower the reprojection error in the cost function. In Fig. 5(a) each method has incorrect poses, with the data-driven approach producing a slightly more accurate 3D pose estimate than the physics-based approach.

VIII. Conclusion

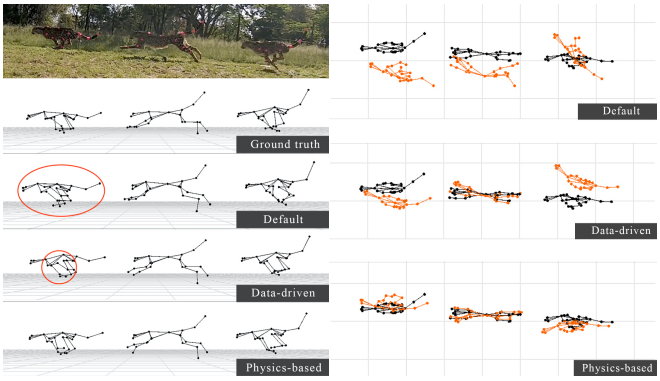
Both the data-driven and physics-based approaches provide significantly better 3D reconstructions as opposed to the default kinematic method. The developed methods provide reasonable 3D reconstructions of the cheetah in the wild using monocular video.

It is not entirely clear which of these methods should be favored as this would depend on the application. The data-driven approach is quick to converge to a solution compared with the physics-based approach. However, the latter provides better dynamic tracking of the cheetah's COM. Both yield similar and reasonably accurate pose estimates that are consistent with MPJPE scores in human research [21].



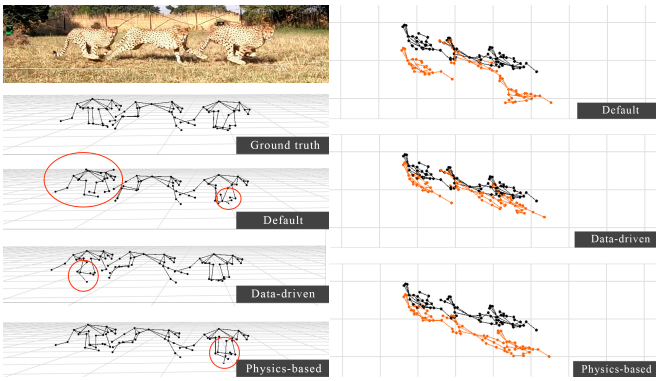
(a) (b)

Fig. 3: Visual comparison of T1 steady-state run for all monocular 3D reconstruction methods. (a) side-view. (b) top-view. The default method is the only one that clearly shows a substantial error in pose.



(a) (b)

Fig. 4: Visual comparison of T1 maneuver for all monocular 3D reconstruction methods. (a) side-view. (b) top-view. The physics-based approach clearly outperforms both in terms of tracking the global dynamics of the cheetah.



(a) (b)

Fig. 5: Visual comparison of T3 maneuver for all monocular 3D reconstruction methods. (a) side-view. (b) top-view. The default method provides substantial body sliding and the data-driven approach appears to provide the most accurate result.

Thus, the physics based approach is best suited for applications focused on biomechanics, whereas the data driven approach is more suited for the analysis of individual pose estimates.

That said, drawbacks for both methods have been identified. The data-driven approach is only as good as the ground truth dataset used to train the models, and in this work, there are potential deficiencies in the size, diversity, and quality of the dataset. On the other hand, the physics-based approach requires the contact timing to be known a priori. This is challenging in wild and uncertain environments as shown in Table II. Although the absence of a reliable 3D ground truth somewhat hindered the evaluation process, the methods developed in this work were found to produce reasonably accurate monocular 3D reconstruction of cheetahs in their natural environment.

Looking ahead, we plan to investigate challenges related to scalability and pre-determined contact states. Both of these issues are prevalent in this subject area, as the cheetah is a fast animal that requires high frame rates to capture, and the environment it traverses is often uneven and unpredictable. This motivates an investigation into other methods that do not suffer from these same issues, namely deep reinforcement learning or sampling-based methods [22], [23].

ACKNOWLEDGMENT

The authors would like to thank Google for their generous support of this research support as well as Ann van Dyk Cheetah Centre (Haartebeespoort, South Africa).

REFERENCES

[1] S. Seok, A. Wang, M. Y. Chuah, D. Otten, J. Lang, and S. Kim, "Design principles for highly efficient quadrupeds and implementation on the mit cheetah robot," in *2013 IEEE International Conference on Robotics and Automation*, 2013, pp. 3307–3312. 1

[2] D. Joska, L. Clark, N. Muramatsu, R. Jericevich, F. Nicolls, A. Mathis, M. W. Mathis, and A. Patel, "AcinoSet: A 3d pose estimation dataset and baseline models for cheetahs in the wild," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 05 2021. [Online]. Available: https://doi.org/10.1109/ICRA48506.2021.9561338 1, 2

[3] Z. da Silva, S. Shield, P. E. Hudson, A. M. Wilson, F. Nicolls, and A. Patel, "Markerless 3d kinematics and force estimation in cheetahs," *Scientific Reports*, vol. 14, no. 1, p. 10579, 2024. 1, 3, 4

[4] H. Sidenbladh, M. J. Black, and D. J. Fleet, "Stochastic Tracking of 3D Human Figures Using 2D Image Motion," in *Computer Vision — ECCV 2000*, D. Vernon, Ed. Berlin, Heidelberg: Springer, 2000, pp. 702–718. 1

[5] N. R. Howe, M. E. Leventon, and W. T. Freeman, "Bayesian reconstruction of 3d human motion from single-camera video," in *Proceedings of the 12th International Conference on Neural Information Processing Systems*, ser. NIPS'99. Cambridge, MA, USA: MIT Press, 1999, p. 820–826. 1, 2

[6] M. E. Leventon and W. T. Freeman, "Bayesian estimation of 3-d human motion," technical report tr 98-06, Mitsubishi electric research labs, Tech. Rep., 1998. 1

[7] J. Shlens, "A tutorial on principal component analysis," *CoRR*, vol. abs/1404.1100, 2014. [Online]. Available: http://arxiv.org/abs/1404.1100 1

[8] R. Bowden, "Learning statistical models of human motion," in *IEEE Workshop on Human Modeling, Analysis and Synthesis, CVPR*, vol. 2000, 2000. 2

[9] L. Tanco and A. Hilton, "Realistic synthesis of novel human movements from a database of motion capture examples," in *Proceedings Workshop on Human Motion*, 2000, pp. 137–142. 2

[10] J. Tilmanne and T. Dutoit, "Expressive Gait Synthesis Using PCA and Gaussian Modeling," in *Motion in Games*, R. Boulic, Y. Chrysanthou, and T. Komura, Eds. Berlin, Heidelberg: Springer, 2010, pp. 363–374. 2

[11] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," 2017. [Online]. Available: https://arxiv.org/abs/1705.03098 2

[12] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3d human pose estimation in the wild using improved cnn supervision," in *2017 international conference on 3D vision (3DV)*. IEEE, 2017, pp. 506–516. 2

[13] A. Gosztolai, S. Günel, V. Lobato-Ríos, M. Pietro Abrate, D. Morales, H. Rhodin, P. Fua, and P. Ramdya, "Liftpose3d, a deep learning-based approach for transforming two-dimensional to three-dimensional poses in laboratory animals," *Nature methods*, vol. 18, no. 8, pp. 975–981, 2021. 2

[14] M. L. Felis, K. Mombaur, and A. Berthoz, "An optimal control approach to reconstruct human gait dynamics from kinematic data," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, 2015, pp. 1044–1051. 2

[15] R. M. Schemschat, D. Clever, M. L. Felis, E. Chiovetto, M. Giese, and K. Mombaur, "Joint torque analysis of push recovery motions during human walking," in *2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, 2016, pp. 133–139. 2

[16] X. Wei and J. Chai, "Videomocap: Modeling physically realistic human motion from monocular video sequences," *ACM Trans. Graph.*, vol. 29, no. 4, jul 2010. [Online]. Available: https://doi.org/10.1145/1778765.1778779 2

[17] Z. Li, J. Sedlar, J. Carpentier, I. Laptev, N. Mansard, and J. Sivic, "Estimating 3d motion and forces of human-object interactions from internet videos," 2021. [Online]. Available: https://arxiv.org/abs/2111.01591 2

[18] D. Rempe, L. J. Guibas, A. Hertzmann, B. Russell, R. Villegas, and J. Yang, "Contact and human dynamics from monocular video," 2020. [Online]. Available: https://arxiv.org/abs/2007.11678 2

[19] E. Gärtner, M. Andriluka, H. Xu, and C. Sminchisescu, "Trajectory optimization for physics-based reconstruction of 3d human pose from monocular video," 2022. [Online]. Available: https://arxiv.org/abs/2205.12292 2

[20] L. Cong, P. Ruppel, Y. Wang, X. Pan, N. Hendrich, and J. Zhang, "Efficient human motion reconstruction from monocular videos with physical consistency loss," in *SIGGRAPH Asia 2023 Conference Papers*, 2023, pp. 1–9. 2

[21] S. Shimada, V. Golyanik, W. Xu, and C. Theobalt, "Physcap: Physically plausible monocular 3d motion capture in real time," *ACM Transactions on Graphics*, vol. 39, no. 6, dec 2020. 2, 6

[22] B. Huang, L. Pan, Y. Yang, J. Ju, and Y. Wang, "Neural mocon: Neural motion control for physically plausible human motion capture," 2022. [Online]. Available: https://arxiv.org/abs/2203.14065 2, 7

[23] Y. Yuan, S.-E. Wei, T. Simon, K. Kitani, and J. Saragih, "Simpoe: Simulated character control for 3d human pose estimation," 2021. [Online]. Available: https://arxiv.org/abs/2104.00683 2, 5, 7

[24] S. Shimada, V. Golyanik, W. Xu, P. Pérez, and C. Theobalt, "Neural monocular 3d human motion capture with physical awareness," *ACM Transactions on Graphics (ToG)*, vol. 40, no. 4, pp. 1–15, 2021. 2

[25] E. Gärtner, M. Andriluka, E. Coumans, and C. Sminchisescu, "Differentiable dynamics for articulated 3d human motion reconstruction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13 190–13 200. 2

[26] P. E. Hudson, S. A. Corr, and A. M. Wilson, "High speed galloping in the cheetah (Acinonyx jubatus) and the racing greyhound (Canis familiaris): spatio-temporal and kinetic characteristics," *Journal of Experimental Biology*, vol. 215, no. 14, pp. 2425–2434, 07 2012. [Online]. Available: https://doi.org/10.1242/jeb.066720 3, 4

[27] N. Muramatsu, Z. da Silva, D. Joska, F. Nicolls, and A. Patel, "Improving 3d markerless pose estimation of animals in the wild using low-cost cameras," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 3770–3776. 3