HybridCap: Inertia-Aid Monocular Capture of Challenging Human Motions

Han Liang\textsuperscript{1}, Yannan He\textsuperscript{1}, Chengfeng Zhao\textsuperscript{1}, Mutian Li\textsuperscript{1}, Jingya Wang\textsuperscript{1,2}, Jingyi Yu\textsuperscript{1,2}, Lan Xu\textsuperscript{1,2}\textsuperscript{*}

\textsuperscript{1}School of Information Science and Technology, ShanghaiTech University
\textsuperscript{2}Shanghai Frontiers Science Center of Human-centered Artificial Intelligence
{lianghan, heyn, zhaochf, limt1, wangjingya, yujingyi, xulan1}@shanghaitech.edu.cn

Abstract

Monocular 3D motion capture (mocap) is beneficial to many applications. The use of a single camera, however, often fails to handle occlusions of different body parts and hence it is limited to capture relatively simple movements. We present a light-weight, hybrid mocap technique called HybridCap that augments the camera with only 4 Inertial Measurement Units (IMUs) in a learning-and-optimization framework. We first employ a weakly-supervised and hierarchical motion inference module based on cooperative pure residual recurrent blocks that serve as limb, body and root trackers as well as an inverse kinematics solver. Our network effectively narrows the search space of plausible motions via coarse-to-fine pose estimation and manages to tackle challenging movements with high efficiency. We further develop a hybrid optimization scheme that combines inertial feedback and visual cues to improve tracking accuracy. Extensive experiments on various datasets demonstrate HybridCap can robustly handle challenging movements ranging from fitness actions to Latin dance. It also achieves real-time performance up to 60 fps with state-of-the-art accuracy.

Introduction

The past ten years have witnessed a rapid development of human motion capture (Davison, Deutscher, and Reid 2001; Hasler et al. 2009; Stoll et al. 2011; Wang et al. 2017), which benefits broad applications like VR/AR, gaming, sports and movies. However, capturing challenging human motions in a light-weight and convenient manner remains unsolved.

The high-end vision-based solutions require attaching dense optical markers (Vicon 2022) or expensive multi-camera setup (Stoll et al. 2011; Joo et al. 2015; Collet et al. 2015; Joo, Simon, and Sheikh 2018) to capture professional motions, which are undesirable for consumer-level usage. Recent learning-based methods enable robust human capture from monocular RGB video (Kanazawa et al. 2019; Kocabas, Athanasiou, and Black 2020; Zheng et al. 2019; Xiang, Joo, and Sheikh 2019). They require specific human templates for space-time coherent capture (Habermann et al. 2019; Xu et al. 2018b, 2020; Habermann et al. 2020), or utilize parametric human model (Kanazawa et al. 2018; Mahmood et al. 2019; Kolotouros et al. 2019; Kocabas, Athanasiou, and Black 2020; Loper et al. 2015). However, these monocular methods are fragile to capture specific challenging motions such as fitness actions or Latin dance, which suffer from complex motion patterns and severe self-occlusion. Recent advances compensate occlusions and challenging motions using probabilistic or attention-based partial occlusion modeling (Kolotouros et al. 2021; Kocabas et al. 2021), or using the generative or weakly-supervised prior (Rempe et al. 2021; He et al. 2021). But the above methods still suffer from the inherent monocular ambiguity due to the lack of reliable observation for self-occluded regions.

In stark contrast, combining occlusion-unaware body-worn sensors for robust motion capture has been widely explored (Henschel, Von Marcard, and Rosenhahn 2020; Pons-Moll et al. 2011; Gilbert et al. 2019; Zhang et al. 2020; Kaufmann et al. 2021). Utilizing Inertial Measurement Units (IMUs) for motion inertia recording is a very popular trend. However, most previous works (Pons-Moll et al. 2010, 2011; Von Marcard, Pons-Moll, and Rosenhahn 2016; Malleson et al. 2017; von Marcard et al. 2018; Gilbert et al. 2019; Malleson, Collomosse, and Hilton 2019; Zhang et al. 2020) utilize multi-view video or a relatively large amount of IMUs (from 8 to 17), which is undesirable for daily usage.
Recently, learning-based approaches (Huang et al. 2018; Yi, Zhou, and Xu 2021) enable a real-time motion capture with only 6 IMUs. But the lack of visual cue leads to inherent drifts and overlay artifacts.

In this paper, we tackle the above challenges and present HybridCap – a high-quality inertia-aid monocular approach for capturing challenging human motions, as shown in Fig. 1. We revisit the light-weight hybrid setting using a single RGB camera and sparse IMUs (only 4 in our setting), and reframe such hybrid motion capture into the neural data-driven realm. It enables practical and robust human motion capture under challenging scenarios at 60 fps with state-of-the-art accuracy.

To this end, we introduce a learning-and-optimization framework, consisting of a hybrid motion inference module and a robust hybrid optimization scheme. The former is formulated in a weakly supervised and hierarchical multi-stage framework. Specifically, we adopt a fully differentiable architecture in an analysis-by-synthesis fashion without explicit 3D ground truth annotation. Thus, for dense hybrid weak supervision, we propose a new largest multimodal human motion dataset called Hybrid Challenging Motions (HCM) consisting of 176-minute challenging motions recorded from 14 RGB cameras and 17 IMUs with a total of 8.9M images.

Our network takes the sequential 2D pose estimation and motion inertia from our light-weight hybrid setting as input, and learns to infer 3D human motion by comparing predicted tracking results against the dense hybrid observations. To handle the challenging motions, we further formulate our network into a series of cooperative subtasks, including a limb tracker, a body tracker, a root tracker and a hybrid IK (inverse kinematics) solver. For all these four subtasks, we also propose a novel bone tracking block design consisting of light-weight pure sequential recurrent layers with skip connection for temporal motion state modeling, which is trained to consider bone length while tracking or solving. Such hierarchical multi-stage design makes full use of our hybrid weak supervision by progressively narrowing the search space of plausible motions in a coarse-to-fine manner, so as to enable efficient and effective inference of challenging motions.

Finally, besides the data-driven motion characteristics from previous module, current multi-modal input also encodes reliable visual and inertial hints, especially for the non-occluded regions. Thus, we further propose a robust hybrid optimization scheme with inferential prior to refine the inferred results. It directly fits 2D and inertial observations to improve the accuracy while keeping multi-modal inferential characteristics to preserve motion plausibility.

To summarize, our main contributions include:

- A real-time and accurate motion capture approach augmenting the light-weight monocular setting using only 4 IMUs aiding, achieving significant superiority to state-of-the-arts.
- A novel network with a hierarchical framework containing cooperative bone tracking blocks with bone length awareness, and a robust optimization scheme combining inferential prior with input observations, improving capture results effectively.
- A new largest multi-modal human motion dataset containing a wide range of challenging motions along with abundant records of RGB cameras and IMUs.

Related Work

Based on our inertia-aid monocular setup, we focus on human motion capture from optical and inertial solutions.

Optical Motion Capture. Marker-based methods (Vicon 2022; Vlasic et al. 2007) have achieved success in capturing professional human motions which are widely utilized in industry, but they are inapplicable for daily usage due to expensive and tedious setup. The exploration of markerless mocap (Bregler and Malik 1998; De Aguiar et al. 2008; Theobalt et al. 2010) has made great progress in order to get rid of body-worn markers and pursue efficiency. Benefiting from researches on parametric human models (Anguelov et al. 2005; Loper et al. 2015; Pavlakos et al. 2019;?) in the last decade, various data-driven approaches are proposed to estimate 3D human pose and shape by optimizing (Huang et al. 2017; Lassner et al. 2017; Bogo et al. 2016; Kolotouros, Pavlakos, and Daniilidis 2019) or directly regressing (Kanazawa et al. 2018, 2019; Kocabas, Athanasiou, and Black 2020; Zanfir et al. 2021) human model parameters. Taking specific template mesh as prior, multi-view (Gall et al. 2010; Stoll et al. 2011; Liu et al. 2013; Robertini et al. 2016; Pavlakos et al. 2017; Simon et al. 2017; Xu et al. 2018a) and monocular (Xu et al. 2018b; Habermann et al. 2019; Xu et al. 2020; Habermann et al. 2020) template-based approaches combine free-form and parametric methods, which produce high quality skeletal and surface motions. Besides, to alleviate the inherent estimation ambiguity from 2D input to 3D motion, recent approaches (Kolotouros et al. 2021; Kocabas et al. 2021; Liang et al. 2023) handle complex patterns using probabilistic or attention-based semantic modeling, (Rempe et al. 2021; He et al. 2021) learn to model generative or weakly-supervised prior to solve unseen and non-periodic motions. However, these methods still suffer from challenging motions, especially for rare pose patterns and extreme self-occlusion.

Inertial Motion Capture. To overcome the limitations of vision cues only, another category of works propose to use IMUs. Previously, purely inertial methods using large amounts of sensors like Xsens MVN (Movella 2022) has been commercially used. However, intrusive capture system prompts researchers forward to sparse-sensor setup. SIP (Von Marcard et al. 2017), which uses only 6 IMUs, presents a pioneering exploration. However, limitations of its traditional optimization framework make real-time application impractical. Recent data-driven works (Huang et al. 2018; Yi, Zhou, and Xu 2021; Yi et al. 2022) achieve great improvements on accuracy and efficiency with sparse sensors, but substantial drift is still unsolved for challenging motions. Preceding sensor-aid solutions propose to combine IMUs with videos (Gilbert et al. 2019; Henschel, Von Marcard, and Rosenhahn 2020; Malleson, Collomosse, and Hilton 2019; Malleson et al. 2017), RGB-D cameras (Hel-
ten et al. 2013; Zheng et al. 2018), optical markers (Andrews et al. 2016) or even LiDAR (Ren et al. 2023). Although these approaches partially solve scene-occlusion problem and correct drift effectively, they are restricted from either undesirable system complexity or implausible estimation for challenging motions.

**Preliminary and Overview**

The goal of our work is to capture challenging 3D human motions using a single camera with few inertial sensors aiding, which suffers from complex motion patterns and extreme self-occlusion. Fig. 2 provides an overview of HybridCap, which relies on a template mesh of the actor and makes full usage of multi-modal input in a learning-and-optimization framework. In the inference stage, our hierarchical design extracts different characteristics from multi-modal observations and learns to estimate plausible motions in a weakly supervised manner. Then, a robust optimization stage is introduced to refine the skeletal motions to increase the tracking accuracy and overlay performance.

**Template and Motion Representation.** We first scan the actor with a 3D body scanner to generate the textured template mesh of the actor. Then, we rig it automatically by fitting the Skinned Multi-Person Linear Model (SMPL) (Loper et al. 2015) to the template mesh and transferring the SMPL skinning weights to our scanned mesh. The kinematic skeleton is parameterized as \( S = [\theta, R, t] \), including the Euler angles \( \theta \in \mathbb{R}^{N_J \times 3} \) of the \( N_J \) joints, the global rotation \( R \in \mathbb{R}^3 \) and translation \( t \in \mathbb{R}^3 \). Furthermore, let \( \Theta \) denotes the 6D representation (Zhou et al. 2019) of the global rotation and joint rotations. Then, we can formulate \( S = M(\Theta, t) \) where \( \mathcal{M} \) denotes the motion transformation between various representations.

**Input Preprocessing.** Our system takes the RGB video, inertial measurements and a well pre-scanned template of the actor as the overall input. Given an image frame, we extract \( N_M \) 2D keypoints \( p \in \mathbb{R}^{N_M \times 2} \) and corresponding confidence \( \sigma \in \mathbb{R}^{N_M} \) using OpenPose (Cao et al. 2017). To generalize to various camera settings during inference time, we refer (Shimada et al. 2021) to use canonicalized 2D keypoints \( p_\text{c} \) by projecting them onto \( Z = 1 \) plane. Then we transform inertial measurements from inertial frame \( F_I \) into camera frame \( F_C \) and obtain IMU accelerations \( A_n \in \mathbb{R}^3 \) and orientations \( R_{b_n} \in \mathbb{R}^{3 \times 3} \) of \( N_I \) corresponding bones \( b_n \) with calibrated \( R_{12C} \) and \( R_{S2B,n} \):

\[
A_n = R_{12C} A_{I,n} \quad (1)
\]

\[
R_{b_n} = R_{12C} R_{b} R_{S2B,n} \quad (2)
\]

where \( R_{12C} \) is the transformation from \( F_I \) to \( F_C \), and \( R_{S2B,n} \) is the transformation from the \( n \)-th IMU sensor \( F_{s_n} \) to \( F_{b_n} \) of its corresponding bone \( b_n \). Besides, to provide prior knowledge on anthropometry, we heuristically calculate \( N_b = 7 \) key bone lengths \( L_k \) (uparm, lowarm, upleg, lowleg, foot, clavicle, and spine) from the rigging skeleton \( S \) and concatenate them into the input. Thus the overall input of the network in a single frame is \( [p_c, \sigma, R_{b_n}, A, L_k] \in \mathbb{R}^{2N_M + N_J + N_I + 3N_b} \).

**Approach**

**Hybrid Motion Inference**

As illustrated in Fig. 3, we adopt a hierarchical multi-stage motion inference scheme. It takes the sequential 2D pose estimation and motion inertia as input and predicts 3D human motions in a weakly supervised manner without explicit 3D ground truth annotation. To tackle challenging motions, we divide the motion inference task into a series of cooperative subtasks. Specifically, we introduce a limb tracker, a body tracker, a root tracker, and a hybrid IK (inverse kinematics) solver, so as to model the hierarchical knowledge of articulated human body structure.

**Pure Residual Recurrent Blocks.** For temporal motion state modeling in these subtasks, we propose a novel effi-
efficient block design using pure recurrent layers with skip connection. In contrast to block design in TransPose (Yi, Zhou, and Xu 2021) adopting both recurrent (LSTM) and fully-connected (FC) layers, we find that such FC layers are the key bottleneck, which encode non-temporal representation and tend to overfit. Thus we simply remove the FC layers and improve the accuracy.

We further introduce skip connection. Our key intuition is that motion inference or position tracking requires less high-level features. We adopt skip connection enabling easier identity function learning to retain more information from the input layer, where the layer close to the output is responsible for adding low-level details. Thus the low-level 2D and inertial features could be selected and passed to the output layer directly, which helps more accurate and detailed motion inference.

**Hierarchical Bone Trackers** In our hierarchical design, our trackers track bones rather than joints. For a bone $b_n$, we track root-relative positions of its two endpoint $J$ with distance constraint using the known bone lengths $L_{b_n}$. Specifically, the limb tracker focuses on accurately tracking four limbs with reliable inertial measurements. Here, the loss of limb tracker is formulated as:

$$\mathcal{L}_{\text{limb}} = \mathcal{L}_{\text{joint}}^{(\text{limb})} + \mathcal{L}_{\text{bone}}^{(\text{limb})},$$

(3)

where $\mathcal{L}_{\text{joint}}^{(\text{limb})}$ is the 3D joint position loss and $\mathcal{L}_{\text{bone}}^{(\text{limb})}$ is the limb bone length loss. Next, the body tracker estimates the rest rigid body parts concatenating the initial input and well-estimated limb positions as the input. Bone length constraints are also utilized to reduce depth ambiguity. Furthermore, the bone orientations help to narrow the search space to the target joint positions. Similar to the limb tracker loss, the body tracker loss is formulated as:

$$\mathcal{L}_{\text{body}} = \mathcal{L}_{\text{joint}}^{(\text{body})} + \mathcal{L}_{\text{bone}}^{(\text{body})},$$

(4)

Note that $\mathcal{L}_{\text{joint}}$ is formulated as the reprojection error to all camera views, which guides the corresponding tracker to predict joint positions in a weakly supervised manner:

$$\mathcal{L}_{\text{joint}} = \sum_{t=1}^{T} \sum_{c=1}^{N_C} \sum_{j=1}^{N_J} \sigma_{c,j}^{(t)} ||\Pi_{c}(\hat{J}_{c,j}^{(t)} + t) - p_{c,j}^{(t)}||^2,$$

(5)

where $\sigma_{c,j}^{(t)}$ denotes the confidence of 2D joint $p_{c,j}^{(t)}$, $\Pi_{c}$ denotes the projection function of camera $c$, $t$ denotes global translation from hybrid optimization of full observations; $N_J$ denotes the number of bone endpoints (8 for limb tracker and 7 for body tracker). Then, the bone length loss is formulated as the $L_2$ loss between the predicted bone length and ground-truth bone length $L_{b_n}$:

$$\mathcal{L}_{\text{bone}} = \sum_{t=1}^{T} \sum_{n=1}^{N_B} \sum_{i=1}^{2} ||\hat{J}_{b_n,i}^{(t)} - \hat{J}_{b_n,i}^{(t)}||^2 - L_{b_n}^2,$$

(6)

where the predicted bone length can be calculated by the distance of two output endpoint positions $\hat{J}_{b_n,i}^{(t)} (i = 0, 1)$ of bone $b_n$. Note that $N_B$ is the number of target bones which is 4 for the limb tracker and 8 for the body tracker.

**Hybrid Inverse Kinematics Solver.** Based on the accurate tracking of bones, we introduce our hybrid IK solver and root tracker to solve rotations and translation respectively. The initial input and well-estimated root-relative 3D joints are concatenated and fed into our hybrid IK solver, which outputs global rotation and local joint rotations $\hat{\Theta}$ in the 6D representation. Then we perform forward kinematics using predicted rotations $\hat{\Theta}$ and bone lengths to obtain refined root-relative 3D joints. Next we send them with the initial input into the root tracker which predicts root position $t$ (i.e. global translation) in the camera frame.

To utilize our dense hybrid weak supervision, we further calculate $N_M$ 3D marker positions $P_m(\hat{\Theta}, t)$ attached to the
skeleton (corresponding to body joints and face landmarks of OpenPose), \( N_I \) bone orientations \( \hat{\mathbf{R}}_{b_n}(\hat{\Theta}) \), and \( N_I \) simulated IMU sensor positions \( \hat{\mathbf{P}}_{n}(\hat{\Theta}, \hat{t}) \) respectively. The IK loss is formulated as:

\[
L_{IK} = \lambda_{2D} L_{2D} + \lambda_{acc} L_{acc} + \lambda_{ori} L_{ori} + \lambda_{prior} L_{prior} + \lambda_{trans} L_{trans}.
\]  

(7)

The 2D reprojection loss \( L_{2D} \) ensures each estimated 3D marker \( \hat{\mathbf{P}}_m \) projects onto the corresponding 2D keypoint \( \mathbf{p}_{e,m} \) in all camera views, formulated as:

\[
L_{2D} = \sum_{t=1}^{T} \sum_{c=1}^{N_C} \sum_{m=1}^{N_M} \sigma_{c,m}^{(t)} \mathcal{D}(\Pi_c(\hat{\mathbf{P}}_{m}^{(t)}(\hat{\Theta}, \hat{t}))) - \mathbf{p}_{e,m}^{(t)} \|_2^2,
\]  

(8)

where \( \sigma_{c,m}^{(t)} \) denotes the confidence of 2D keypoint \( \mathbf{p}_{c,m} \) and \( \Pi_c \) is the projection function of camera \( c \). Then, we introduce the acceleration loss \( L_{acc} \) to encourage the network to learn the implicit physical constraints and generate plausible motions:

\[
L_{acc} = \sum_{t=2}^{T-1} \sum_{c=1}^{N_C} \| \hat{\mathbf{A}}_n^{(t)}(\hat{\Theta}, \hat{t}) - \mathbf{A}_n^{(t)} \|_2^2,
\]  

(9)

where \( \hat{\mathbf{A}}_n^{(t)} \) is the estimated acceleration calculated from predicted IMU position \( \hat{\mathbf{P}}_n \), formulated as below where \( st \) is the sampling time:

\[
\hat{\mathbf{A}}_n^{(t)} = (\hat{\mathbf{P}}_n^{(t+1)} - 2\hat{\mathbf{P}}_n^{(t)} + \hat{\mathbf{P}}_n^{(t-1)})/st^2.
\]  

(10)

We further propose the orientation loss \( L_{ori} \), which ensures the predicted orientation \( \hat{\mathbf{R}}_{b_n} \) of each bone bound with IMU sensor fits the observation \( \hat{\mathbf{R}}_{b_n} \) of the corresponding IMU measurement:

\[
L_{ori} = \sum_{t=1}^{T-1} \sum_{n=1}^{N_I} \| \hat{\mathbf{R}}_{b_n}^{(t)}(\hat{\Theta}) - \mathbf{R}_{b_n} \|_2^2.
\]  

(11)

Hybrid Motion Optimization

Despite the data-driven motion inference stage learns the mapping from multi-modal observations to 3D motions, the generalization error is non-negligible due to noisy 2D detection and inertial measurements.

We further introduce a hybrid motion optimization stage to refine the skeletal motions to increase the tracking accuracy and overlay performance. It jointly utilizes the learned 3D prior from the network of multi-modal weak supervision, the 2D keypoints in the visible regions as well as inertial measurements.

In this stage, we first initialize the skeletal motion sequence \( S \) using network output by representation transformation \( \mathcal{M}(\hat{\Theta}, \hat{t}) \) and then perform the optimization procedure. We adopt the Euler angle representation so that the joint angles \( \theta \) of \( S \) locate in the pre-defined range \([\theta_{\text{min}}, \theta_{\text{max}}]\) of physically plausible joint angles to prevent unnatural poses. Our energy function is formulated as:

\[
E(S) = \lambda_{3D} E_{3D} + \lambda_{2D} E_{2D} + \lambda_{acc} E_{acc} + \lambda_{ori} E_{ori}.
\]  

(12)

Here, \( E_{3D} \) enforces the final motion sequence close to the predicted one; \( E_{2D} \) ensures that each final 3D marker reprojects onto the corresponding 2D keypoint. Besides, we adopt the acceleration energy \( E_{acc} \) to enforce the final motion to be temporally consistent with the network estimating accelerations \( \hat{\mathbf{A}}_n \) supervised by \( N_I \) IMUs and the measured ground-truth accelerations \( \mathbf{A}_n \) from \( N_I \) input IMUs. Specifically, the acceleration term \( E_{acc} \) is formulated as:

\[
E_{acc} = \sum_{t=2}^{T-1} \sum_{n=1}^{N_I} \gamma^{(t)} \| \mathbf{A}_n^{(t)}(S) - \hat{\mathbf{A}}_n^{(t)} \|_2^2
\]  

\[
+ \sum_{t=2}^{T-1} \sum_{n=N_I+1}^{N_I} \| \mathbf{A}_n^{(t)}(S) - \hat{\mathbf{A}}_n^{(t)} \|_2^2,
\]  

(13)

where the first term means that we directly use the acceleration observation from the input IMUs and the second term...
Experiments

We run our pipeline on a PC with an i7-10700k CPU and RTX3070 GPU, where the inference module and optimization module take 2.8(±0.4) ms and 6.2(±3.5) ms respectively, achieving 60 fps. In this section, first, we describe the datasets used for training and evaluation. Next, we further qualitatively (Fig. 5) and quantitatively (Tab. 1) illustrate that our method outperforms previous state-of-the-arts. We also provide extensive ablation studies (Tab. 2, Tab. 3 and Fig. 6, Fig. 7) to evaluate our technical design and input setting. Finally, we present more qualitative results in Fig. 4.

Ablation Study

Evaluation on inference module. We first evaluate our inference module by comparing to the variants of our approach and previous state-of-the-arts. The quantitative results are provided in Rows 1-5 in Tab. 2. VIBE structure (Row 1) shows VIBE variant with 4 IMUs and RGB input, where we concatenate inertial input and key bone lengths with image features obtained from pre-trained feature extractor used in the vanilla VIBE. TransPose structure (Row 2) shows TransPose variant with 4 IMUs and RGB input, where we concatenate 2D keypoints and key bone lengths with inertial
input. We train both two networks using the same training set and losses as ours. The results show our network with nuanced designs to progressively fuse multi-modal input is superior to both two previous state-of-the-arts. Row 3-5 corresponds to the variants without our entire inference module, without using the skip connection in our block design, without using bone length, respectively. **Purely recurrent design (Row 4)** without FC layers (adopted in Row 2), effectively alleviates overfitting. **Skip connection** (Row 4) operation enables to preserve more low-level features from the input layer, and the **bone length** (Row 5) awareness effectively narrows the search space, which are the two key designs to improve the inference results. Corresponding qualitative results are provided in Fig. 6 (a), which demonstrates our two nuanced designs also contribute to improving the overlay performance. These results illustrate the effectiveness of inference module and highlight the contribution of our algorithmic component designs.

**Evaluation on optimization module.** To evaluate the effectiveness of our optimization module design, we further compare to the two variants without the entire optimization module and without using the inertial terms in Eqn. 13 and Eqn. 14, respectively. The quantitative results are provided in Rows 6-7 of Tab. 2. Our inertial terms improve position accuracy while preserving motion plausibility (acceleration) from the inference module. The qualitative results provided in Fig. 6 (b) show our robust optimization improves the overlay performance effectively. Note that these variants suffer from tracking loss especially for the limbs with fast and challenging motions (generalization error). In contrast, our full pipeline achieves more robust capture.

![Figure 6: Qualitative evaluation on our inference module and optimization module.](image)

| Ablation                      | MPJPE ↓ | PCK↑ | AE↓ |
|-------------------------------|---------|------|-----|
| 1) VIBE net                  | 75.4    | 76.5 | 34.2|
| 2) TransPose net             | 61.0    | 82.7 | 19.2|
| 3) w/o inference net         | 136.5   | 51.1 | 82.1|
| 4) w/o skip connection       | 54.5    | 86.9 | 18.7|
| 5) w/o bone length           | 65.8    | 81.3 | 19.6|
| 6) w/o optimization          | 57.5    | 85.3 | 18.6|
| 7) w/o inertial terms        | 51.2    | 87.4 | 62.2|
| 8) Ours (complete)           | **43.3**| **90.1**| **17.9**|

Table 2: Quantitative evaluation on our network structure, bone length utilization and optimization configurations.

![Figure 7: Qualitative evaluation on our multi-modal input setting and IMU number configurations.](image)

| Input setting   | MPJPE ↓ | PCK↑ | AE↓ |
|-----------------|---------|------|-----|
| a) 4 IMUs Only  | 104.7   | 57.4 | 18.5|
| b) RGB Only     | 77.6    | 74.2 | 28.1|
| c) 2-IMU-aid    | 65.0    | 81.0 | 23.4|
| d) **4-IMU-aid**| **43.3**| **90.1**| **17.9**|
| e) 12-IMU-aid   | 36.1    | 93.2 | 15.1|

Table 3: Quantitative evaluation on our input configuration.

**Evaluation on 4-IMU-aid setting.** Finally, to verify the necessity and rationality of our multi-modal input setting with only 4 IMUs aiding, we compare several variants of our approach with various network input settings. As shown in Tab. 3, even using 2 IMUs aiding, our multi-modal input outperforms pure IMU or RGB input under single modality. Besides, our approach with 4 IMUs significantly outperforms the one with only two IMUs (one for left arm and one for right leg) and closes to the one with tedious 12 IMUs, which serves as a good compromise of acceptable performance and light-weight convenient capture setting. As shown in Fig. 6, our inertia-aid setting augments the camera by alleviating its inherent defects in terms of depth ambiguity and occlusion. As shown in Fig. 7, 4-IMU-aid setting enables to constantly track frequently occluded extremities while minimizing the intrusive body-worn sensors.

**Conclusion**

We present a practical approach HybridCap to capture challenging 3D human motions using only a single camera and 4 IMUs and achieve superior results compared to previous state-of-the-art methods. HybridCap uses a learning-and-optimization framework through novel cooperative bone tracking blocks with bone length awareness and an optimization-based refinement module with inferential prior. We also propose a new largest multi-modal human motion dataset, called HCM dataset, to evaluate their approach thoroughly. The experimental results demonstrate the robustness of HybridCap in capturing challenging human motions in various scenarios. We believe that it is a significant step for convenient and robust capture of human motions, with many potential applications in VR/AR and motion evaluations for gymnastics and dancing.
Acknowledgements

This work was supported by Shanghai YangFan Program (21YF1429500), Shanghai Local college capacity building program (22010502800), NSFC programs (61976138, 61977047), the National Key Research and Development Program (2018YFB2100500), STCSM (2015F0203-000-06), and SHMEC (2019-01-07-00-01-E00003).

References

Agarwal, S.; Mierle, K.; and The Ceres Solver Team. 2022. Ceres Solver. https://github.com/cess-solver/cess-solver. Accessed: 2023-03-26.

Andrews, S.; Huerta, I.; Komura, T.; Sigal, L.; and Mitchell, K. 2016. Real-time physics-based motion capture with sparse sensors. In Proceedings of the 13th European conference on visual media production (CVMP 2016), 1–10.

Anguelov, D.; Srinivason, P.; Koller, D.; Thrun, S.; Rodgers, J.; and Davis, J. 2005. SCAPE: Shape Completion and Animation of People. In ACM SIGGRAPH 2005 Papers, SIGGRAPH ’05, 408–416. New York, NY, USA: Association for Computing Machinery. ISBN 9781450378253.

Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; and Black, M. J. 2016. Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., Computer Vision – ECCV 2016, 561–578. Cham: Springer International Publishing.

Bregler, C.; and Malik, J. 1998. Tracking people with twists and exponential maps. In Computer Vision and Pattern Recognition (CVPR).

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In Computer Vision and Pattern Recognition (CVPR).

Collet, A.; Chuang, M.; Sweeney, P.; Gillett, D.; Evseev, D.; Calabrese, D.; Hoppe, H.; Kirk, A.; and Sullivan, S. 2015. High-quality streamable free-viewpoint video. ACM Transactions on Graphics (TOG), 34(4): 69.

Davison, A. J.; Deutscher, J.; and Reid, I. D. 2001. Markerless Motion Capture of Complex Full-body Movement for Character Animation. In Eurographics Workshop on Computer Animation and Simulation.

De Aguiar, E.; Stoll, C.; Theobalt, C.; Ahmed, N.; Seidel, H.-P.; and Thrun, S. 2008. Performance capture from sparse multi-view video. In ACM SIGGRAPH 2008 papers, 1–10.

Gall, J.; Rosenhahn, B.; Brox, T.; and Seidel, H.-P. 2010. Optimization and Filtering for Human Motion Capture. International Journal of Computer Vision (IJCV), 87(1–2): 75–92.

Gilbert, A.; Trumble, M.; Malleson, C.; Hilton, A.; and Collomosse, J. 2019. Fusing visual and inertial sensors with semantics for 3d human pose estimation. International Journal of Computer Vision, 127(4): 381–397.

Habermann, M.; Xu, W.; Zollhöfer, M.; Pons-Moll, G.; and Theobalt, C. 2019. LiveCap: Real-Time Human Performance Capture From Monocular Video. ACM Transactions on Graphics (TOG), 38(2): 14:1–14:17.

Habermann, M.; Xu, W.; Zollhöfer, M.; Pons-Moll, G.; and Theobalt, C. 2020. DeepCap: Monocular Human Performance Capture Using Weak Supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Hasler, N.; Rosenhahn, B.; Thormahlen, T.; Wand, M.; Gall, J.; and Seidel, H.-P. 2009. Markerless motion capture with unsynchronized moving cameras. In Computer Vision and Pattern Recognition (CVPR), 224–231.

He, Y.; Pang, A.; Chen, X.; Liang, H.; Wu, M.; Ma, Y.; and Xu, L. 2021. Challencap: Monocular 3d capture of challenging human performances using multi-modal references. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11400–11411.

Helten, T.; Muller, M.; Seidel, H.-P.; and Theobalt, C. 2013. Real-time body tracking with one depth camera and inertial sensors. In Proceedings of the IEEE international conference on computer vision, 1105–1112.

Henschel, R.; Von Marcard, T.; and Rosenhahn, B. 2020. Accurate long-term multiple people tracking using video and body-worn IMUs. IEEE Transactions on Image Processing, 29: 8476–8489.

Huang, Y.; Bogo, F.; Lassner, C.; Kanazawa, A.; Gehler, P. V.; Romero, J.; Akhter, I.; and Black, M. J. 2017. Towards Accurate Marker-Less Human Shape and Pose Estimation over Time. In 2017 International Conference on 3D Vision (3DV), 421–430.

Huang, Y.; Kaufmann, M.; Aksan, E.; Black, M. J.; Hilliges, O.; and Pons-Moll, G. 2018. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. ACM Transactions on Graphics (TOG), 37(6): 1–15.

Joo, H.; Liu, H.; Tan, L.; Gui, L.; Nabbe, B.; Matthews, I.; Kanade, T.; Nobuhara, S.; and Sheikh, Y. 2015. Panoptic Studio: A Massively Multiview System for Social Motion Capture. In Proceedings of the IEEE International Conference on Computer Vision, 3334–3342.

Joo, H.; Simon, T.; and Sheikh, Y. 2018. Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end Recovery of Human Shape and Pose. In Computer Vision and Pattern Recognition (CVPR).

Kanazawa, A.; Zhang, J. Y.; Felsen, P.; and Malik, J. 2019. Learning 3D Human Dynamics From Video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Kaufmann, M.; Zhao, Y.; Tang, C.; Tao, L.; Twigg, C.; Song, J.; Wang, R.; and Hilliges, O. 2021. EM-POSE: 3D Human Pose Estimation From Sparse Electromagnetic Trackers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 11510–11520.

Kocabas, M.; Athanasiou, N.; and Black, M. J. 2020. VIBE: Video Inference for Human Body Pose and Shape Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
Kocabas, M.; Huang, C.-H. P.; Hilliges, O.; and Black, M. J. 2021. PARE: Part Attention Regressor for 3D Human Body Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 11127–11137.

Kolotouros, N.; Pavlakos, G.; Black, M. J.; and Daniilidis, K. 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2252–2261.

Kolotouros, N.; Pavlakos, G.; and Daniilidis, K. 2021. Probabilistic Modeling for Human Mesh Recovery. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 11605–11614.

Lassner, C.; Romero, J.; Kiefel, M.; Bogo, F.; Black, M. J.; and Gehler, P. V. 2017. Unite the people: Closing the loop between 3d and 2d human representations. In Proceedings of the IEEE conference on computer vision and pattern recognition, 6050–6059.

Li, R.; Yang, S.; Ross, D. A.; and Kanazawa, A. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 13401–13412.

Liang, H.; Zhang, W.; Li, W.; Yu, J.; and Xu, L. 2023. InterGen: Diffusion-based Multi-human Motion Generation under Complex Interactions. arXiv preprint arXiv:2304.05684.

Liu, Y.; Gall, J.; Stoll, C.; Dai, Q.; Seidel, H.-P.; and Theobalt, C. 2013. Markerless motion capture of multiple characters using multiview image segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35(11): 2720–2735.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A Skinned Multi-person Linear Model. ACM Trans. Graph., 34(6): 248:1–248:16.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of Motion Capture As Surface Shapes. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Malleson, C.; Collomosse, J.; and Hilton, A. 2019. Real-time multi-person motion capture from multi-view video and IMUs. International Journal of Computer Vision, 1–18.

Malleson, C.; Gilbert, A.; Trumble, M.; Collomosse, J.; Hilton, A.; and Volino, M. 2017. Real-time full-body motion capture from video and imus. In 2017 International Conference on 3D Vision (3DV), 449–457. IEEE.

Movella. 2022. Movella Xsens Products. https://www.movella.com/products/xsens. Accessed: 2023-03-26.

Pavlakos, G.; Zhou, X.; Derpanis, K. G.; and Daniilidis, K. 2017. Harvesting Multiple Views for Marker-less Human Pose Annotations. In Computer Vision and Pattern Recognition (CVPR).

Pons-Moll, G.; Baak, A.; Gall, J.; Leal-Taixe, L.; Mueller, M.; Seidel, H.-P.; and Rosenhahn, B. 2011. Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In 2011 International Conference on Computer Vision, 1243–1250. IEEE.

Pons-Moll, G.; Baak, A.; Helten, T.; Müller, M.; Seidel, H.-P.; and Rosenhahn, B. 2010. Multisensor-fusion for 3d full-body human motion capture. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 663–670. IEEE.

Rempe, D.; Birdal, T.; Hertzmann, A.; Yang, J.; Sridhar, S.; and Guibas, L. J. 2021. HuMoR: 3D Human Motion Model for Robust Pose Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 11488–11499.

Ren, Y.; Zhao, C.; He, Y.; Cong, P.; Liang, H.; Yu, J.; Xu, L.; and Ma, Y. 2023. LiDAR-aid Inertial Poser: Large-scale Human Motion Capture by Sparse Inertial and LiDAR Sensors. IEEE Transactions on Visualization and Computer Graphics, 1–11.

Robertini, N.; Casas, D.; Rhodin, H.; Seidel, H.-P.; and Theobalt, C. 2016. Model-based Outdoor Performance Capture. In International Conference on 3D Vision (3DV).

Shimada, S.; Golyanik, V.; Xu, W.; Pérez, P.; and Theobalt, C. 2021. Neural Monocular 3D Human Motion Capture with Physical Awareness. ACM Transactions on Graphics, 40(4).

Simon, T.; Joo, H.; Matthews, I.; and Sheikh, Y. 2017. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In Computer Vision and Pattern Recognition (CVPR).

Stoll, C.; Hasler, N.; Gall, J.; Seidel, H.-P.; and Theobalt, C. 2011. Fast articulated motion tracking using a sums of Gaussians body model. In International Conference on Computer Vision (ICCV).

Theobalt, C.; de Aguiar, E.; Stoll, C.; Seidel, H.-P.; and Thrun, S. 2010. Performance capture from multi-view video. In Image and Geometry Processing for 3-D Cinematography, 127–149. Springer.

Tsuchida, S.; Fukayama, S.; Hamasaki, M.; and Goto, M. 2019. AIST Dance Video Database: Multi-genre, Multi-dancer, and Multi-camera Database for Dance Information Processing. In Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, 501–510. Delft, Netherlands.

Vicon. 2022. Vicon Mocap Systems. https://www.vicon.com. Accessed: 2023-03-26.

Vlasic, D.; Adelsberger, R.; Vannucci, G.; Barnwell, J.; Gross, M.; Matusik, W.; and Popović, J. 2007. Practical motion capture in everyday surroundings. ACM transactions on graphics (TOG), 26(3): 35–es.

von Marcard, T.; Henschel, R.; Black, M.; Rosenhahn, B.; and Pons-Moll, G. 2018. Recovering Accurate 3D Human
Pose in The Wild Using IMUs and a Moving Camera. In *European Conference on Computer Vision (ECCV)*.

Von Marcard, T.; Pons-Moll, G.; and Rosenhahn, B. 2016. Human pose estimation from video and imus. *IEEE transactions on pattern analysis and machine intelligence*, 38(8): 1533–1547.

Von Marcard, T.; Rosenhahn, B.; Black, M. J.; and Pons-Moll, G. 2017. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer Graphics Forum*, volume 36, 349–360. Wiley Online Library.

Wang, Y.; Liu, Y.; Tong, X.; Dai, Q.; and Tan, P. 2017. Outdoor Markerless Motion Capture With Sparse Handheld Video Cameras. *Transactions on Visualization and Computer Graphics (TVCG)*.

Xiang, D.; Joo, H.; and Sheikh, Y. 2019. Monocular Total Capture: Posing Face, Body, and Hands in the Wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, L.; Liu, Y.; Cheng, W.; Guo, K.; Zhou, G.; Dai, Q.; and Fang, L. 2018a. FlyCap: Markerless Motion Capture Using Multiple Autonomous Flying Cameras. *IEEE Transactions on Visualization and Computer Graphics*, 24(8): 2284–2297.

Xu, L.; Xu, W.; Golyanik, V.; Habermann, M.; Fang, L.; and Theobalt, C. 2020. Eventcap: Monocular 3d capture of high-speed human motions using an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4968–4978.

Xu, W.; Chatterjee, A.; Zollhöfer, M.; Rhodin, H.; Mehta, D.; Seidel, H.-P.; and Theobalt, C. 2018b. MonoPerfCap: Human Performance Capture From Monocular Video. *ACM Transactions on Graphics (TOG)*, 37(2): 27:1–27:15.

Yi, X.; Zhou, Y.; Habermann, M.; Shimada, S.; Golyanik, V.; Theobalt, C.; and Xu, F. 2022. Physical Inertial Poser (PIP): Physics-aware Real-time Human Motion Tracking from Sparse Inertial Sensors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yi, X.; Zhou, Y.; and Xu, F. 2021. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG)*, 40(4): 1–13.

Zanfir, A.; Bazavan, E. G.; Zanfir, M.; Freeman, W. T.; Sukthankar, R.; and Sminchisescu, C. 2021. Neural descent for visual 3d human pose and shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14484–14493.

Zhang, Z.; Wang, C.; Qin, W.; and Zeng, W. 2020. Fusing wearable imus with multi-view images for human pose estimation: A geometric approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2200–2209.

Zheng, Z.; Yu, T.; Li, H.; Guo, K.; Dai, Q.; Fang, L.; and Liu, Y. 2018. HybridFusion: Real-time Performance Capture Using a Single Depth Sensor and Sparse IMUs. In *European Conference on Computer Vision (ECCV)*.