Markerless Motion Tracking with Noisy Video and IMU Data

Soyong Shin, Zhixiong Li, and Eni Halilaj

***Abstract*— *Objective*: Marker-based motion capture, considered the gold standard in human motion analysis, is expensive and requires trained personnel. Advances in inertial sensing and computer vision offer new opportunities to obtain research-grade assessments in clinics and natural environments. A challenge that discourages clinical adoption, however, is the need for careful sensor-to-body alignment, which slows the data collection process in clinics and is prone to errors when patients take the sensors home. *Methods*: We propose deep learning models to estimate human movement with noisy data from videos (VideoNet), inertial sensors (IMUNet), and a combination of the two (FusionNet), obviating the need for careful calibration. The video and inertial sensing data used to train the models were generated synthetically from a marker-based motion capture dataset of a broad range of activities and augmented to account for sensor-misplacement and camera-occlusion errors. The models were tested using real data that included walking, jogging, squatting, sit-to-stand, and other activities. *Results*: Compared to state-of-the-art models, IMUNet was as accurate, while VideoNet and FusionNet reduced mean (± std) root-mean-squared errors by 7.6 ± 5.4° and 5.9 ± 3.3°, respectively, on calibrated data. Importantly, these models were less sensitive to noise than existing approaches, reducing errors by up to 14.0 ± 5.3° for sensor-misplacement errors of up to 30.0 ± 13.7° and by up to 7.4 ± 5.5° for joint-center-estimation errors of up to 101.1 ± 11.2 mm, across joints. *Conclusion*: These tools offer clinicians and patients the opportunity to estimate movement with research-grade accuracy, without the need for time-consuming calibration steps or the high costs associated with trusted commercial products such as Theia3D or Xsens, helping democratize the diagnosis, prognosis, and treatment of musculoskeletal conditions.**

***Index Terms*— Biomechanics, Computer vision, Deep learning, Human motion capture, Inertial measurement units**

I. Introduction

MARKERLESS motion tracking could transform rehabilitation monitoring [6], [35], athletic performance [37], and biomechanics research at large [36]. A key challenge toward better understanding of human movement, both in health and pathology, has been the limited set of tools available to study it. Marker-based motion capture, the most widely used approach, is limited to research laboratories and specialty clinics due to equipment cost and required expertise. Inertial measurement units (IMUs) and computer vision algorithms are now accessible alternatives that could democratize gait analysis for researchers, clinicians, and patients everywhere. Such tools will strengthen the feedback loop between research and clinical practice, enabling clinics to collect research-grade data for tailored care and researchers to receive large datasets for increased statistical power and more accurate machine-learning models. Currently, inertial sensing requires careful sensor-to-body calibration and suffers from integration drift [14], [28], while vision-based approaches require a relatively high number of cameras for optimal accuracy [16], resulting in prohibitive data storage demands.

The need for sensor-to-body calibration makes inertial sensors less practical for clinical use and remote monitoring, especially when patients are required to mount the sensors themselves. Filtering [2], [8], [22] and physics-based approaches [1], [5], [38], which rely on a calibration step, are therefore better suited for laboratory studies. Deep learning models trained on augmented data can account for sensor misplacement, making them better suited for remote monitoring, but until now they have been limited to walking and running since the large datasets used to train them include only those activities [12], [24], [30]. The Archive of Motion Capture as Surface Shapes (AMASS) dataset [23], however, includes a wider array of human movements and could improve the generalizability of these deep learning models. It is the largest publicly available motion capture database, amassing data from 23 laboratories and 495 subjects.

Vision-based motion tracking is affected by both the number and quality of videos, in addition to relying on spatial calibration. A typical approach for extracting three-dimensional (3-D) kinematics from video data is to extract two-dimensional (2-D) joint centers from individual videos using a neural network [7], [29] and then fit a parametric statistical mesh of the human body on the 2-D or triangulated 3-D joint centers using top-down optimization [4], [10], [26] (Fig. 1). While systematic analyses of how the accuracy of kinematics changes with video data quality are currently missing, a commercial software of growing popularity (Theia Markerless, Kingston, Ontario, Canada) reports accuracies of 2.6 to 13.2° with 8 cameras across lower-extremity joints [18]. This camera density may not be feasible in clinical settings or patient homes, motivating the need for high accuracies with fewer cameras. A recently

This work was supported in part by a Korean Government Fellowship, a Center for Machine Learning and Health Fellowship from Carnegie Mellon University, and U.S. National Science Foundation Award CBET 2145473.

Soyong Shin is with the Carnegie Mellon University, Pittsburgh, PA, USA (e-mail: soyongs@andrew.cmu.edu).

Zhixiong Li is with the Carnegie Mellon University, Pittsburgh, PA, USA (e-mail: zhixionl@andrew.cmu.edu).

Eni Halilaj is with the Carnegie Mellon University, Pittsburgh, PA, USA (e-mail: ehalilaj@andrew.cmu.edu).

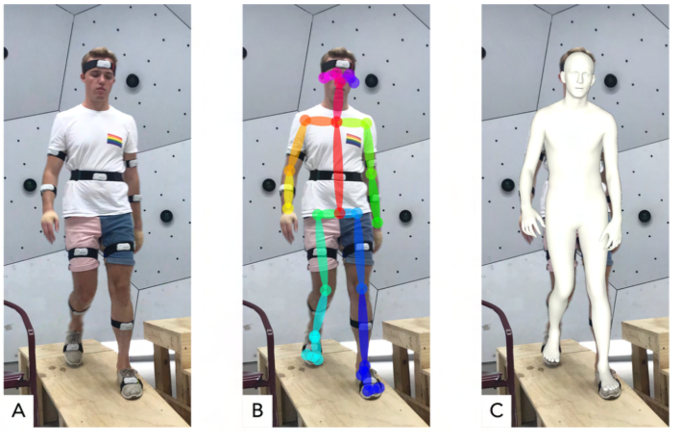

Fig. 1. Meshed Model Fitting. (A) Original image. (B) Joint centers are estimated through single-view or multi-view computer vision models (convolutional neural networks, such as OpenPose). (C) A statistical meshed model is fitted to the given joint centers via top-down optimization, using a cost function that minimizes the distance between the joint centers in the image and the joint centers of the meshed model, while bound to satisfy physiological constraints.

released open-source package (OpenCap) can estimate lower-body kinematics with a mean absolute error of 4.5° using two cameras, but generalizability to data from laboratories outside the one that collected the training data remains to be demonstrated [42]. Fusion of video and inertial sensing data via unconstrained optimization [10], [44], biomechanically constrained optimization [27], or neural networks [19], [40], [48] has been proposed as an alternative for improving accuracy, with initial success. Yet, priorly proposed approaches continue to require careful sensor-to-body calibration steps.

Here we present deep neural networks for the estimation of joint kinematics directly from noisy inertial and video data, obviating the need for careful calibration steps. These models support applications where the adoption of gait analysis hinges on data acquisition speed. In this paper, we (1) benchmark these models against state-of-the-art markerless motion tracking methods, (2) characterize model accuracy as a function of data quality, and (3) provide access to all the associated data and code required to reproduce and build on the presented work. We hypothesized that the accuracy of the models proposed here would match that of state-of-the-art models when the data are well-calibrated. Importantly, we hypothesized that the accuracy of these models would surpass that of state-of-the-art models by clinically significant margins ($>$ 5 degrees) when the data are noisy, confirming their suitability for clinical translation.

II. Methods

A. Datasets

To train the neural networks, we used AMASS [23], a publicly available dataset (Table 1). With 495 subjects (214 females and 281 males) performing a diverse set of activities, for a total of 60 hours, AMASS is currently the largest standardized motion capture dataset, making it ideal for model training (Fig. 2A). Represented activities include walking, jogging, dancing, and squatting. To standardize kinematics across markersets used by different laboratories, statistical meshes of the human body parameterized in terms of body pose and shape [20], [26] have been fit to both the AMASS data [9], [21]. The statistical meshed model is parameterized in terms of 72 rotational degrees of freedom and 10 body-shape modes of variation. Given that the AMASS dataset does not include video and inertial data, we generated these data synthetically (Fig. 2A).

Synthetic joint centers were obtained from the body mesh using a mesh-to-joints regression matrix associated with this parametric mesh. To mimic the prediction error of computer vision models, we added three different types of noise to the synthetic joint centers: jittering, high-peak noise, and bias. Frame-to-frame jittering of joint centers was modeled as Gaussian noise (σ = 12 mm). Low-frequency and high-peak noise were modeled as Gaussian noise (σ = 250 mm) with masking (p = 0.03), which accounts for large prediction errors in computer vision models, such as missing joint centers in certain frames due to occlusion. In addition to these two sources of noise, bias due to disagreements between the joint centers of the body meshed model and the joint centers identified in images with neural networks is also often present. We modeled this disagreement as a bias with a magnitude of approximately 25 mm based on empirical evidence. The noise was modeled independently for each joint center because these errors are not consistent across joints. For example, joint centers such as the hips and knees, for which annotators exhibit higher ambiguity during annotation of common datasets, have larger jittering and bias than other joints [31]. Distal joints like the ankle and wrist, however, have larger peak noise since these joints are more likely to be occluded [50]. These three components resulted in a mean noise of 41 ± 30 mm across joints, which is larger than the root-mean-squared error (RMSE) associated with state-of-the-art approaches utilizing four videos [3], [15], [48], providing more robustness for our models.

Synthetic angular velocities and linear accelerations were extracted from 13 virtual IMUs attached to different segments of the body mesh. The calibration and placement of each virtual sensor was based on a sample subject from the Total Capture dataset, which contained real IMU data (Fig. 3). To account for sensor placement errors, we augmented the data by randomly rotating IMUs, with rotations drawn from a Gaussian distribution (σ = 13.5°). Given the orientations and placements of the virtual sensors, we generated angular velocities and linear accelerations by taking numerical derivatives.

To validate and test the models, we used both a publicly available benchmark dataset (Total Capture [41], Test 1) and data collected in our laboratory (Test 2). The mix of test data was intentional, to demonstrate generalizability across different environments and beyond standard benchmark datasets. Total Capture, with five subjects, contains a total of 37 minutes of marker-based motion capture data collected simultaneously with data from 4 RGB cameras and 13 inertial sensors (Movella, Henderson, NV), which were calibrated both spatially and temporally (Fig. 2B). Represented activities include walking, range of motion, acting, and freestyle motion. In our laboratory, we collected data from three healthy adults

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2023.3275775

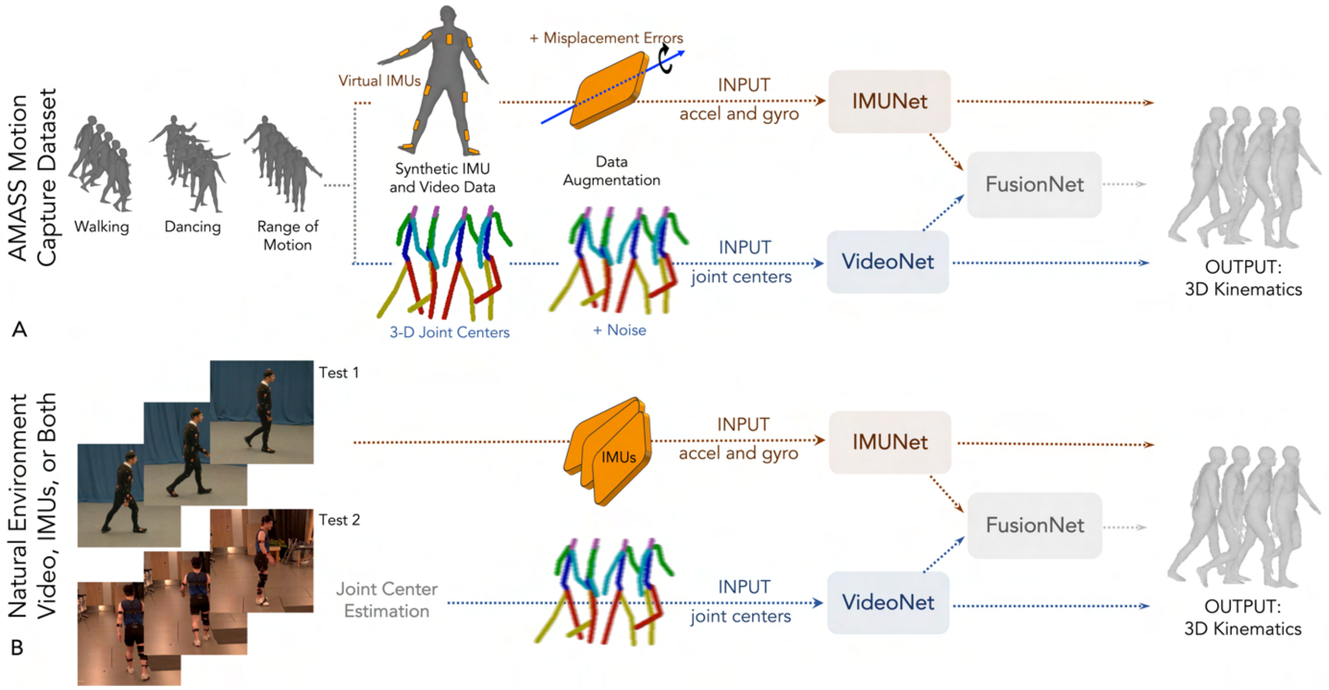

Fig. 2. Overview of the Approach. (A) Synthetic video and IMU data were generated from the AMASS dataset (n = 495 subjects, t = 60 hours), a publicly available marker-based motion capture dataset amassed from 23 different laboratories. Synthetic video data were augmented to account for jittering and occlusion errors, while synthetic IMU data were augmented to account for miscalibration errors. Three models were trained to allow flexibility in data collection: IMUNet, VideoNet, and FusionNet. (B) The models were tested using real data from four videos and 13 IMUs.

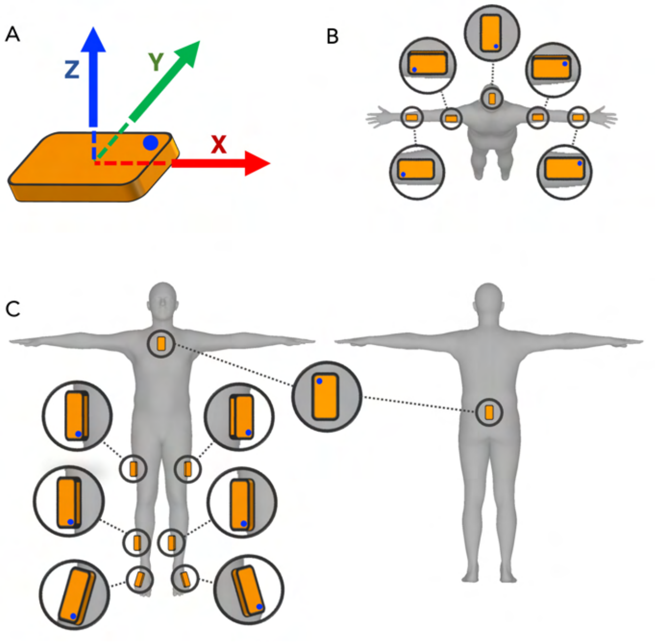

Fig. 3. Virtual IMU Placement. Virtual IMUs were placed on 13 body segments. (A) The IMU orientation (i.e, internal coordinate system relative to a world coordinate system) was derived from marker-based motion tracking data. (B) The top view, illustrating IMU placement on the head and arms. (C) Front and rear views, illustrating the positioning of the back and lower limb sensors.

TABLE I
TRAIN, VALIDATION, AND TEST DATASETS

Dataset	Number of subejcts	Activity type	Total time (minutes)
Train (AMASS)	495 (214 F, 281 M)	Walking, Running, Jogging, Crouch walking, Crawling, Dancing, Squatting, Jumping, etc.,	3731
Validation (Total Capture)	3 (1 F, 2 M)	Walking, Range of motion, Acting, Freestyle motion	14
Test 1 (Total Capture)	5 (1 F, 4 M)	Walking, Range of motion, Acting, Freestyle motion	23
Test 2 (Our data)	3 (3 M)	Squatting, Sit-to-stand, Drop jump, Step up, Walking, Braced walking	17

performing walking, jogging, sit-to-stand, squatting, drop-jump, and step-up tasks, for a total of 17 minutes of data. Walking and jogging were performed while subjects walked normally and while wearing a brace on the knee of the dominant leg to emulate stiff-knee and/or asymmetric gait, which is typical in orthopedics patients. Subjects were equipped with 41 markers (modified Rizzoli markerset) and 13 IMUs (Movella, Henderson, NV). We used 20 infrared cameras (Optitrak, Corvallis, OR) to collect ground truth motion from markers and 4 RGB cameras (Optitrak, Corvallis, OR) to collect video data for the computer vision approaches.

B. Neural Networks

We trained three sets of deep neural networks: IMUNet, VideoNet, and FusionNet. Training of FusionNet consisted of three stages: unimodal encoding, fusion, and kinematics

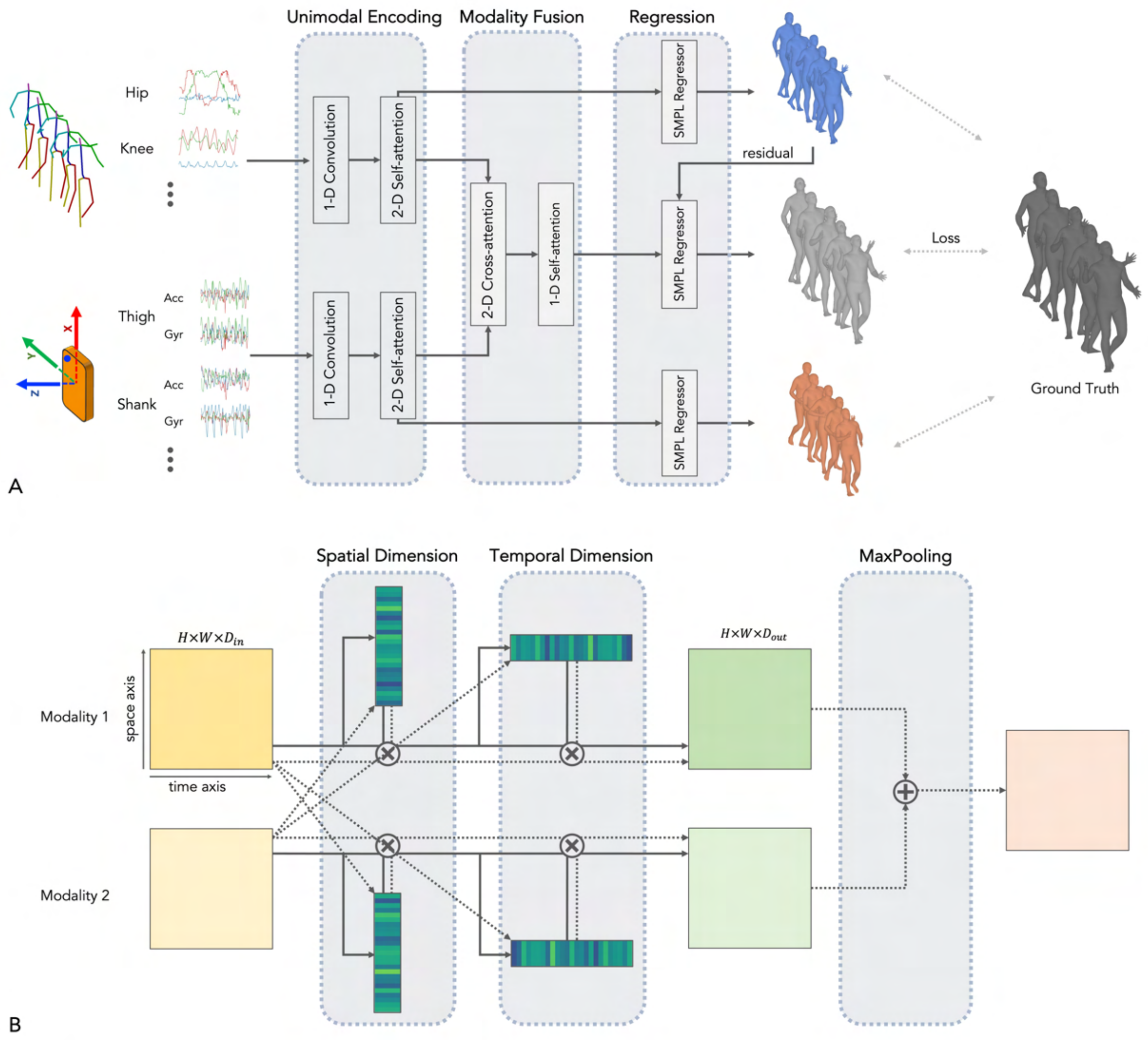

Fig. 4. Neural Network Overview. (A) 3-D joint centers and IMU data were input into 1-D convolutional layers and 2-D self-attention blocks to extract video and IMU features. Features were then fused by a 2-D cross-attention block before being regressed through a 1-D self-attention block and linear layers. A regressor for the FusionNet computed the residual from the output of VideoNet. (B) A 2-D attention mechanism was applied to the feature map in both the space and time axes. Self-attention within each modality is noted by a rigid line, while cross-modality-fusion attention is noted by dotted lines. The cross-attention operation was applied to the video and IMU features across each of the axes. Last, the aggregated feature map was computed by the maximum pooling of the two feature maps.

regression (Fig. 4A), whereas the other two did not include the fusion stage. The unimodal encoding stage was used to extract features from each modality without interaction between modalities. The fusion stage merged these features. The last stage, regression, predicted kinematics from the merged features. For IMUNet and VideoNet, features extracted in the unimodal encoding stage were input directly through the regression stage.

First, we used one-dimensional (1-D) convolutional blocks consisting of three layers, where each layer is followed by an activation function (ReLU) [47] and a regularization function (BatchNorm) [13]. The sizes of the convolutional kernels were set to 1, 3, and 5 to allow each element to refer to various ranges of neighbors. The features were then input into 2-D transformers, which here model the relations within and between the spatial and temporal spectra of the data (Fig. 4B) [43], [49]. For each attention mechanism, we used 4 layers with 8 attention heads and a hidden dimension of 32. In the fusion stage, a cross-attention block was used to compute correlations between the video and inertial data and to align them in both spatial and temporal dimensions. The features were aggregated into a single feature using 1-D max pooling (Fig. 4B). Subsequently, a 1-D transformer collected the temporal information and re-aligned the aggregated features in the time dimension. Finally, three regression networks composed of linear layers were used to predict the body mesh parameters from the unimodal and merged features. Here, we used Gaussian Error Linear Units (GELU) for the activation function to allow higher non-linearity on the prediction [11] and dropout for regularization [39]. In FusionNet, kinematics were predicted through a residual architecture [17], allowing networks to fine-tune the prediction of the unimodal network in a top-down manner. In this architecture, the estimation of i-th iteration was obtained as follows:

$$\hat{\theta}_i = f([\hat{\theta}_{i-1}; y_{fusion}]) + \hat{\theta}_{i-1}. \tag{1}$$

where $f(\cdot)$ is regressor network, y_{fusion} is the fused feature, and $[\cdot;\cdot]$ denotes a concatenation operation. The output was initialized using the proposal from VideoNet, after determining that iterative regression with initialization from a unimodal network converged faster during training and performed better during validation.

The loss functions used to train the models included three terms: pose loss, joint-centers loss, and body-shape loss. Pose loss, l_θ, was defined as the sum of the L-2 distances between ground-truth, θ, and the predicted joint angles $\hat{\theta}$. Joint center position loss, l_p, was defined as the 3-D Euclidean distance between ground truth, p, and predicted, $\hat{p}$, joint centers. Body-shape loss was defined as the L-2 distance between ground truth, β, and predicted shape parameters, $\hat{\beta}$.

$$l_\theta = \frac{1}{N_f N_s} \sum_{t=1}^{N_f} \sum_{s=1}^{N_s} \Omega\left(\theta_{s,t},\ \hat{\theta}_{s,t}\right) \tag{2}$$

$$l_p = \frac{1}{N_f N_j} \sum_{t=1}^{N_f} \sum_{j=1}^{N_j} ||p_{j,t} - \hat{p}_{j,t}|| \tag{3}$$

$$l_\beta = \frac{1}{N_f} \sum_{t=1}^{N_f} \left|\left|\beta_t - \hat{\beta}_t\right|\right| \tag{4}$$

Here, N_f, N_s, N_j denote the number of time frames, body segments associated with IMUs, and joints, respectively, while t, s, j denote the indices for time, body segment, and joint. $\Omega(\theta_1, \theta_2)$ is the magnitude of the relative rotation between θ_1 and θ_2.

C. Model Performance

To test the first hypothesis that our three models match state-of-the-art approaches when the data are well-calibrated, we benchmarked IMUNet, VideoNet, and FusionNet against Xsens [33], SMPLify3D (a 3-D extension of SMPLify) [4], [26], and unconstrained optimization [10], respectively. For primary hypothesis testing, we used the Total Capture data (Test 1). The primary evaluation metric was the mean joint angle RMSE for the lower (knees, hips, and ankles) and upper (shoulders, elbows, and neck) extremity joints, across all the activities. We used paired t tests with Bonferroni correction for multiple comparisons to determine statistically significant differences ($p = 0.013$). Pairing at the joint and degree-of-freedom level was used for all the comparisons. We report all the data as means and standard deviations, after testing for normality using the Anderson-Darling test. Two sets of tests were carried out for additional insight. We split the data by activity (walking, range of motion, acting, and freestyle motion) and joint type (upper and lower extremities) to gain insight into the performance of the models across these different contexts. Repeated measures (RM) analyses of variance (ANOVA) were used to test for difference across activities. Unpaired t tests were used to compare model performance in the upper versus lower extremity joints.

Last, we compared our models to Theia3D, a commercial software of growing popularity [18], using data collected in our laboratory (Test 2). Our VideoNet and FusionNet used data from 4 cameras, while Theia3D requires a minimum of 6 videos. In addition to comparing our models against Theia3D using a RM ANOVA, we also tested that our models' performance is not lower on stiff-knee (braced) gait compared to normal (unbraced) gait using a paired t test.

D. Sensitivity Analysis

To test the second hypothesis, that the models presented here outperform existing approaches when the video and IMU data are noisy, we augmented the calibrated data from Total Capture (Test 1) to replicate real-world errors. Miscalibration errors ranged from 0 to 30.0 ± 13.7° for IMU data, which encompasses the majority of errors encountered in field studies [25], [32]. For video data, we added joint-center errors ranging from 52.6 ± 6.8 to 101.1 ± 11.2 mm, since this range is representative of errors associated with computer vision models [15], [34]. We repeated the experiments 30 times for each noise level to capture varying combinations of misplacement error across body segments. To determine if the performance varied across noise levels and models, we performed a two-way RM ANOVA. The effects of IMU and video noise were modeled in separate statistical analyses. We selected four discrete noise profiles for each analysis.

III. Results

A. Model Performance with Calibrated Data

The three proposed networks matched or outperformed existing approaches when the data were carefully calibrated. IMUNet matched the accuracy of Xsens (8.8 ± 3.4° vs. 8.2 ± 3.3°, $p = 0.3192$), whereas VideoNet and FusionNet outperformed state-of-the-art approaches by a mean ± std of 7.6 ± 5.4° (7.0 ± 3.0° vs. 14.6 ± 7.7°, $p = 0.0013$) and 5.9 ± 3.3° (6.2 ± 2.8° vs. 12.2 ± 5.6°, $p = 0.0002$), respectively (Fig. 5). Across activities and lower-extremity joints, the proposed models performed similarly with Theia3D, despite the latter using six videos when our models used four (Fig. 6). Theia3D achieved an RMSE of 5.6 ± 0.7°, while IMUNet, VideoNet, and FusionNet achieved comparable RMSEs of 6.0 ± 0.9° ($p = 0.4813$), 5.6 ± 0.6° ($p = 0.9195$), and 5.1 ± 0.7 ($p = 0.3940$), respectively.

All the proposed models performed better in lower than upper extremity joints (Fig. 5B), better for standard activities than freestyle motion (Fig. 5C), and similarly between normal and "impaired gait". RMSEs were lower in lower-extremity than upper-extremity joints for IMUNet (6.6 ± 1.2° vs. 11.6 ± 3.1°, $p = 0.0045$), VideoNet (5.2 ± 0.3 vs. 9.2 ± 3.3°, $p = 0.0075$), and FusionNet (4.5 ± 0.3° vs. 8.3 ± 3.0°, $p = 0.0054$). Compared to freestyle motion, walking, range of motion, and acting activities were estimated with a mean ± std RMSE that was at least 4.1 ± 4.0° lower for IMUNet ($p = 0.0060$), 4.1 ± 4.2° lower VideoNet ($p = 0.0089$), and 4.2 ± 3.8 lower for FusionNet ($p = 0.0069$). In comparison to unbraced walking, IMUNet, VideoNet, and FusionNet estimated braced walking kinematics with a mean ± std RMSE that differed

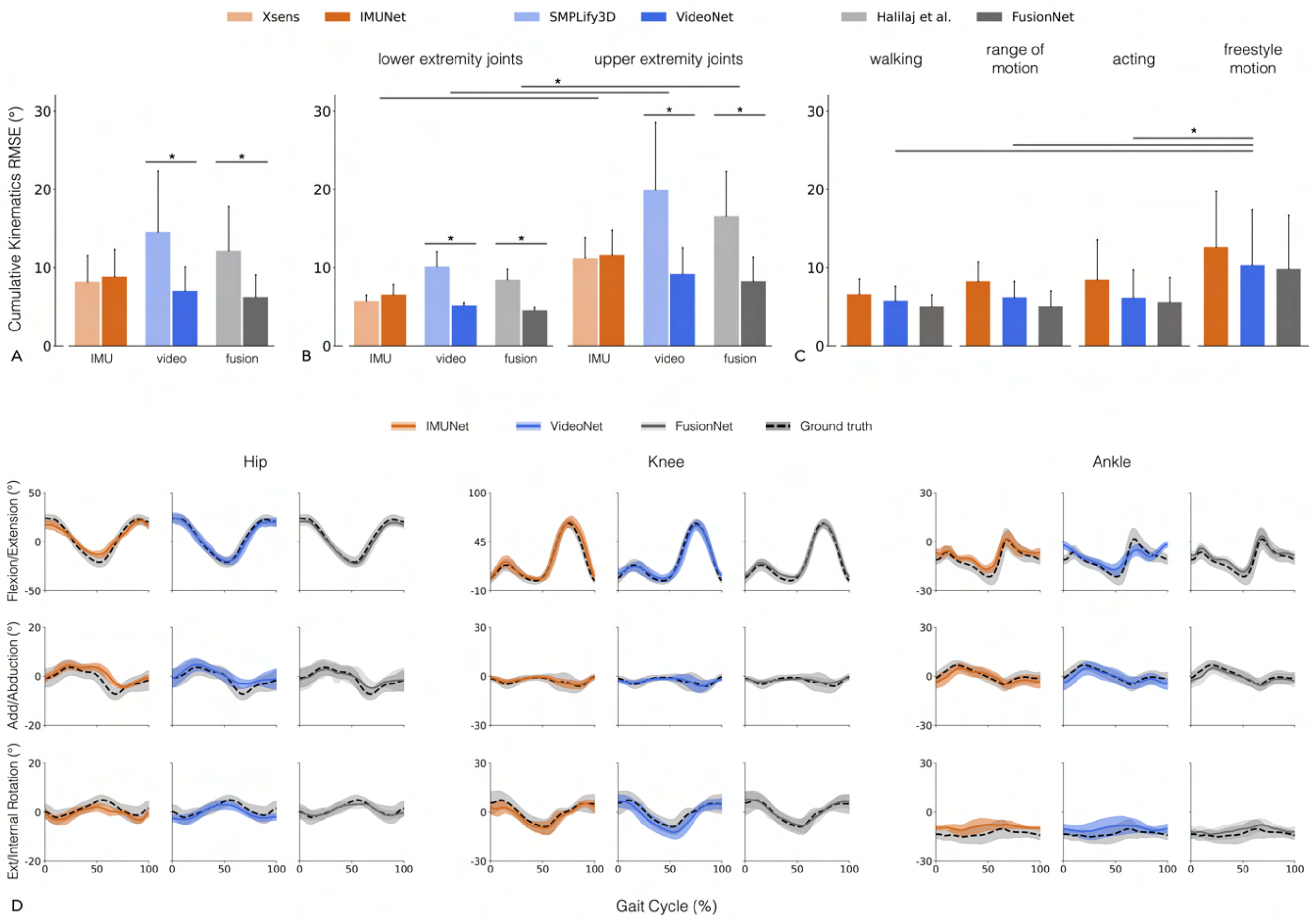

Fig. 5. Model Performance: Total Capture (Test 1). (A) The proposed models matched or outperformed state-of-the-art methods, across activities and joints. (B) All models performed better when estimating lower-extremity joint angles than upper-extremity ones. (C) All models performed better for walking, range of motion, and acting than freestyle motion, which includes fewer cyclical movements. (D) Walking kinematics across the gait cycle (mean and standard deviations) for the lower extremity joints, illustrating that our models outperformed existing markerless approach and that FusionNet performs better than IMUNet and VideoNet.

insignificantly by 0.6 ± 0.5° ($p = 0.0260$), 0.6 ± 0.4° ($p = 0.0398$), and 0.5 ± 0.4° ($p = 0.0348$), respectively.

Fusion of IMU and video data was marginally more accurate than using a single data modality. Across all joints and activities, FusionNet outperformed IMUNet and VideoNet by a mean ± std of 2.6 ± 1.3° (6.2 ± 2.8° vs. 8.8 ± 3.4°, $p = 0.0001$) and 0.8 ± 0.3° (6.2 ± 2.8° vs. 7.0 ± 3.0°, $p < 0.0001$) respectively (Fig. 5D). This result generalized across other activities and joints.

B. Sensitivity to Noise

IMUNet and FusionNet outperformed existing approaches when the IMU data were noisy (Fig. 7A). When IMUs were miscalibrated by up to 30.0 ± 13.7°, IMUNet outperformed Xsens by up to a mean ± std of 14.0 ± 5.3° (13.9 ± 5.1° vs. 27.9 ± 6.2°, $p < 0.0001$), across joints. When IMUs were miscalibrated by 15.0 ± 6.3°, IMUNet outperformed Xsens by a mean ± std of 5.9 ± 3.8° (9.8 ± 3.9° vs. 15.7 ± 4.9°, $p = 0.0006$). FusionNet performed better than both IMUNet and Xsens across noise levels, with larger effects as miscalibration increased. Generally, kinematics predicted via Xsens explained less of the variability in ground-truth kinematics ($R^2 < 0.33$) compared to IMUNet ($R^2 > 0.76$) and FusionNet ($R^2 > 0.90$) (Fig. 8).

VideoNet and FusionNet outperformed existing approaches when the video data were noisy (Fig. 7B). Given joint-center errors of up to 101.1 ± 11.2 mm, VideoNet outperformed SMPLify3D by up to 7.4 ± 5.5° (7.8 ± 2.2° vs. 15.2 ± 7.5°, $p = 0.0017$), across joints. Given joint-center errors of up to 70.5 ± 7.7 mm, VideoNet surpassed SMPLify3D by a mean ± std of 7.1 ± 6.0° (6.5 ± 1.8° vs. 13.6 ± 7.6°, $p = 0.0036$). FusionNet performed better than both IMUNet and SMPLify3D across noise levels, with larger effects as the noise level increased. Generally, kinematics predicted via SMPLify3D explained less of the variability in ground-truth kinematics ($R^2 < 0.55$) than VideoNet ($R^2 > 0.85$) and FusionNet ($R^2 > 0.90$) (Fig. 8).

C. Synthetic Data Quality

The synthetic data used to train the models were generally representative of true inertial sensing and video-based joint centers and the performance of the proposed models degraded only minimally when tested on real data, despite the models being trained using synthetic data. The root-mean-squared

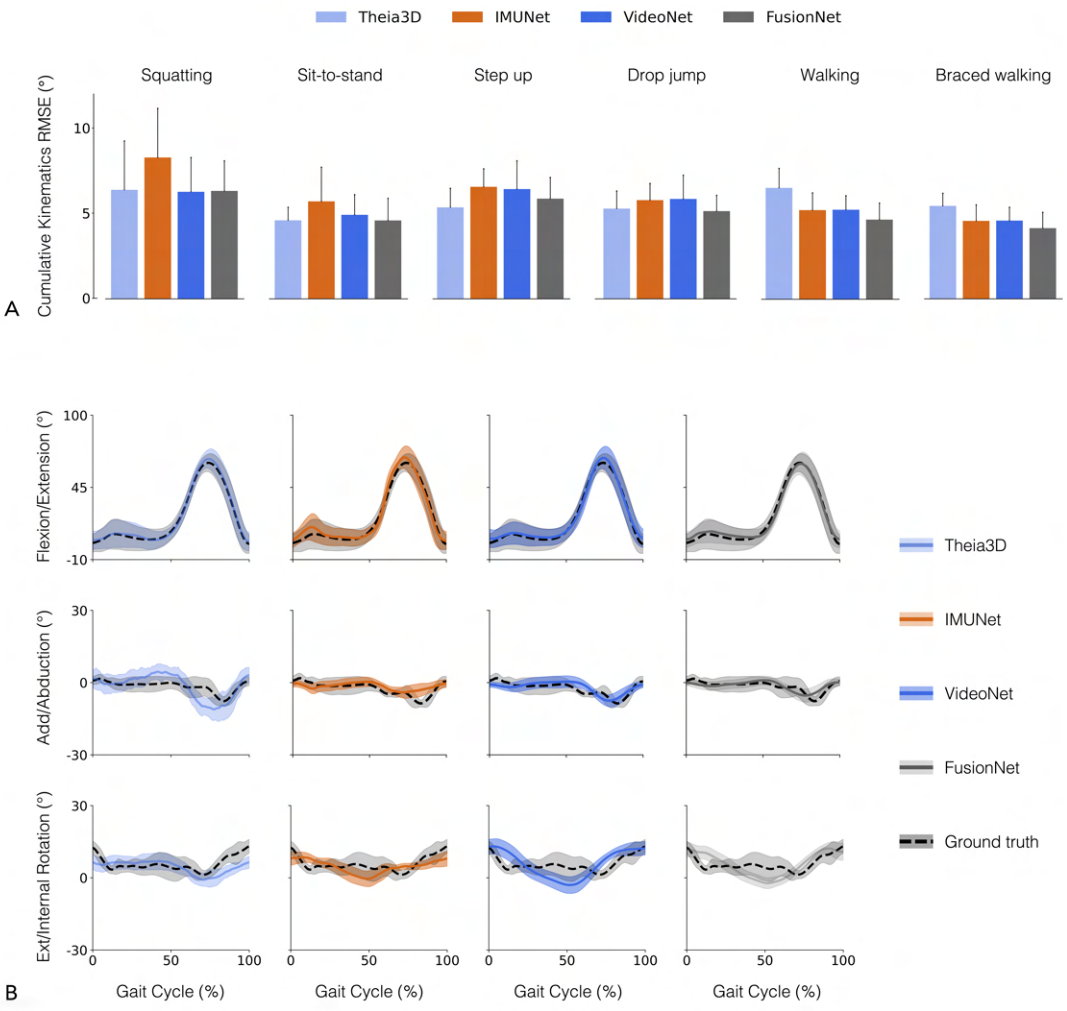

Fig. 6. Model Performance: Our Data (Test 2). (A) Across the lower extremities and activities, the proposed models matched or outperformed the commercial software Theia3D, although Theia3D used six cameras, while VideoNet and FusionNet used four. (B) Walking kinematics (mean and standard deviations) for three degrees of freedom in the knee, illustrating that our models matched or outperformed Theia3D.

deviation (RMSD) of the synthetic joint centers from true joint centers was 78.7 ± 6.7 mm across all the joints, which is greater than the artificial noise we added to the ground truth joint centers in efforts of matching the error of true joint centers (52.6 ± 6.8 mm). The RMSDs of the synthetic inertial data from true inertial data were 30.4°/s (4% of range) for angular velocity and 3.7 m/s^2 (12% of range) for linear acceleration, across all the IMUs. The accuracy of IMUNet, VideoNet, and FusionNet receded by 0.6 ± 0.6°, 0.4 ± 0.2°, and 0.4 ± 0.3°, respectively, when these models were tested on real, instead of synthetic, data (Table 2).

TABLE II
PERFORMANCE WITH REAL AND SYNTHETIC DATA

Models	Input data type	Angle RMSE (°) (mean ± std)
IMUNet	real IMU syn. IMU	5.6 ± 0.7 4.9 ± 0.6
VideoNet	real video syn. video	5.3 ± 0.6 5.0 ± 0.6
FusionNet	real IMU & video syn. IMU & video	4.5 ± 0.4 4.1 ± 0.6

IV. Discussion

The goal of this study was to enable fast motion tracking with inertial sensors and portable cameras, obviating the need for careful sensor-to-body calibration in applications where data acquisition speed is important. We took advantage of synthetic data generation and augmentation approaches to build a large database of inertial and video data from ground truth motion capture and trained deep learning models to predict joint kinematics. Results indicate that the models proposed here outperform state-of-the-art models and are less affected by data quality. Given that we have made the data and code publicly available, these models can be retrained to accommodate data collection with a wide range of sensor densities and locations, depending on the kinematic outcomes of interest. They can be combined with transfer learning to enable better generalization to clinical populations.

When using these models and interpreting the reported findings, a few characteristics of the study must be noted. First, the AMASS dataset used to train the models is limited to healthy adults without major gait pathologies. While it

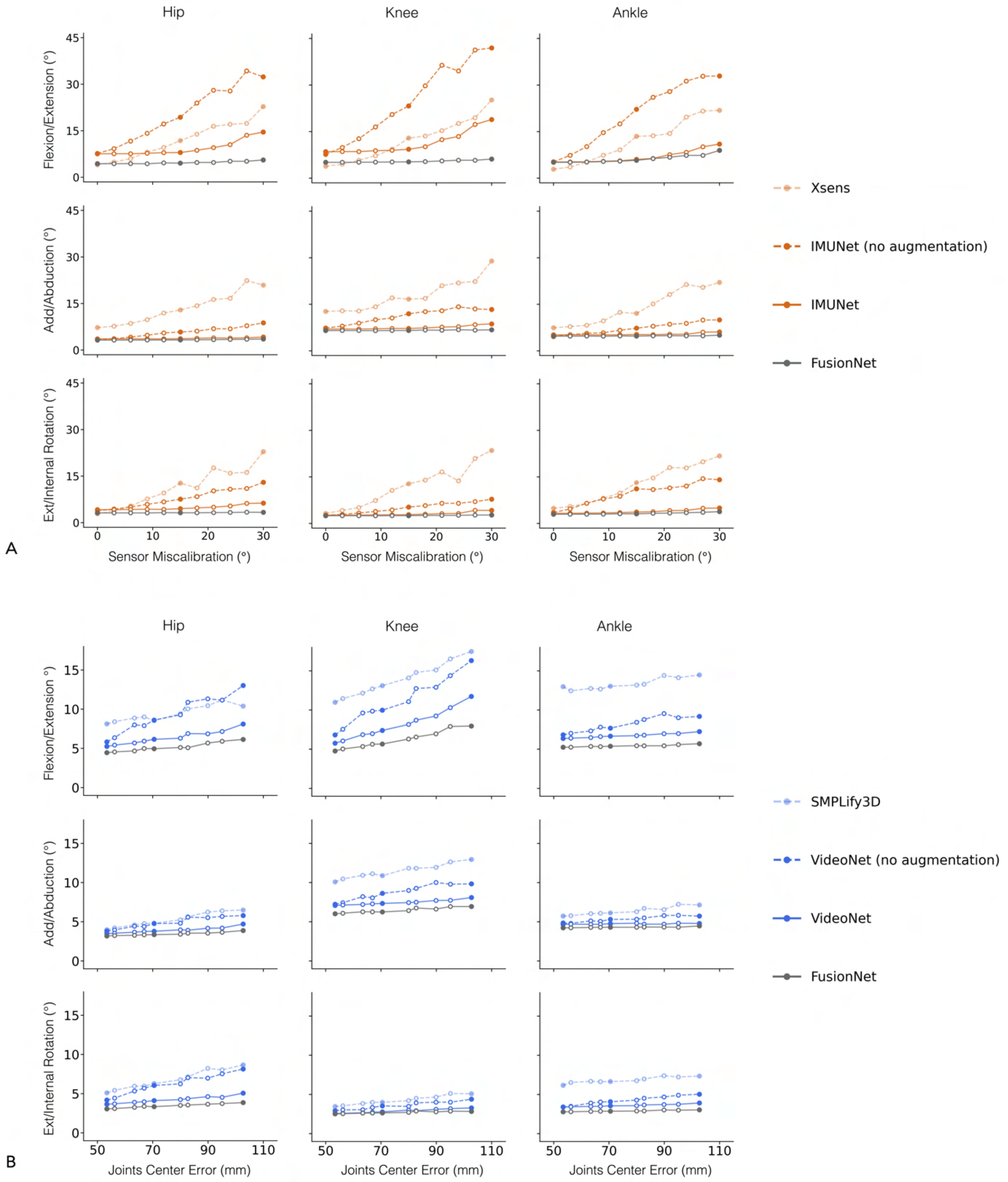

Fig. 7. Model Sensitivity to Sensor and Video Noise. The proposed models were evaluated on data with different levels of noise. (A) IMUNet and FusionNet outperformed Xsens when the IMU-miscalibration error ranged from 0° (baseline) to 30.0 ± 13.7°. (B) VideoNet and FusionNet outperformed SMPLify3D when the joint-center error ranged from 52.6 mm (baseline) to 101.1 ± 11.2 mm.

will be important to either test or finetune the models using data from a population of interest, our preliminary proof-of-concept testing of the models with ”impaired gait” (i.e., braced gait) indicates that the models generalize to gaits not present in the training data. Transfer learning may be used as an alternative strategy toward better generalizability should our results not uphold across populations and types of gaits. Second, while these models may be useful for rapid data

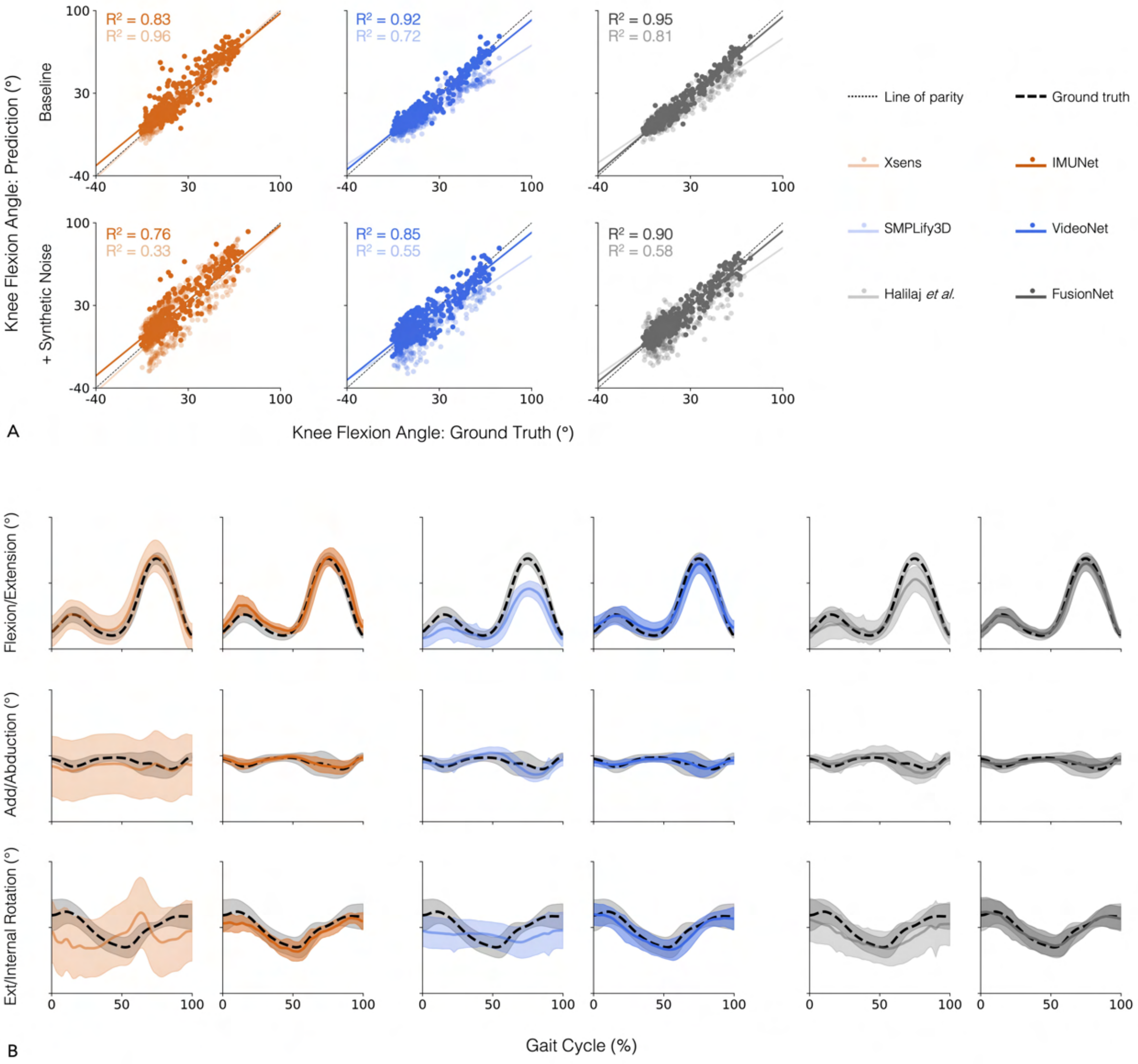

Fig. 8. Model Sensitivity to a Single Noise Profile. The proposed models generally outperformed state-of-the-art approaches both when the sensors were well calibrated and under the synthetic noise condition (i.e., 15.0 ± 6.6° IMU-miscalibration error and 70.5 ± 7.7 mm of joint-center error). A) Parity plots indicate that the proposed models performed better, especially when that data are noisy. The results, presented here only for knee flexion/extension, held across joints, degrees of freedom, and activities. B) Joint angle curves (mean and standard deviation) predicted by IMUNet, VideNet, and FusionNet were not only more accurate, but also less variable, than those predicted by state-of-the-art methods.

collection in clinics and patient homes, we do not expect them to generalize to applications that require higher accuracy. One of the limitations of the AMASS dataset is that the markersets used by the computer vision and graphics communities are not as detailed as those recommended by the International Society of Biomechanics [45], [46]. A third limitation is that, although the models were tested with real IMU and video data (Fig. 2), sensitivity to noise was tested using synthetically added noise (Fig. 7). It remains to be determined whether the observed robustness to different levels of noise will translate to real-world data. This is especially important given that we modeled noise using Gaussian distributions. A final note is that, while we obviated the need for spatial calibration, FusionNet still requires time syncing of the video and IMU data, which may require a hop trial or another salient activity that is easily identifiable by both modalities.

The success of these models is primarily due to the AMASS dataset, spanning a wide array of human movements captured across multiple laboratories. Prior deep learning models for kinematics estimation from IMUs have been limited to the lower extremity and only walking and running activities and have not included tests with real data [24], [30]. Here, we trained models entirely using synthetic data and tested their generalizability using real data. Our analysis revealed general agreement between real and synthetic data, reinforcing the utility of synthetic data. The fact that the performance of these open-source models matches that of expansively licensed commercial tools (Theia Markerless, Kingston, Ontario, Canada) is remarkable.

All the models performed better in estimating lower than upper extremity kinematics, likely due to the more constrained physics that characterizes gait compared to upper extremity motion. Walking, for example, follows passive dynamics principles that neural networks can learn more easily than they can learn arm motion. Similar logic could be used to explain the lower performance of the models during freestyle motion than other activities. Walking, for instance, is cyclical and can be roughly modeled using passive pendular dynamics principles. Freestyle motion, however, is more variable and not typically modeled using simplistic dynamical models that capture most of the relevant physics.

While neural networks were able to learn typical gait patterns and harness data augmentation techniques to minimize sensitivity to noise compared to state-of-the-art approaches, they were not able to harness the complementary nature of IMU and video data well. Although FusionNet maintained relatively stable errors across IMU or video data noise profiles and was generally more accurate than IMUNet and VideoNet, these improvements were not always clinically meaningful (Fig. 7). For example, sagittal plane kinematics in the hip and knee benefit from IMU-video fusion when either the IMU or video data are very noisy, but other degrees of freedom do not. It is therefore important to consider the clinical application when weighing the cost and benefit of additional sensors. The comprehensive data presented here should facilitate these choices. Further, they serve as a baseline for what purely data-driven approaches can achieve as new fusion approaches, including those that rely on biomechanical modeling, are developed.

V. Conclusion

As markerless motion tracking evolves and becomes more pervasive in movement sciences and clinical research, a wide array of tools will enable better matches between the required accuracy for a given application and data collection speed and complexity. In this study, we proposed three neural network models to predict 3-D human kinematics from videos, IMUs, and their fusion. The proposed models performed accurately with noisy data, obviating the need for careful sensor-to-body calibration and a dense number of cameras. These lightweight models enable rapid gait analysis in clinical settings, where the adoption of motion-tracking technology is lagging behind, despite the transformative impact it could have on patient treatment. We have open-sourced the models, code, and synthetic data to allow others to build on this work.

References

[1] Sheldon Andrews, Ivan Huerta, Taku Komura, Leonid Sigal, and Kenny Mitchell. Real-time physics-based motion capture with sparse sensors. In *Proceedings of the 13th European Conference on Visual Media Production (CVMP 2016)*, CVMP 2016, New York, NY, USA, 2016. Association for Computing Machinery.

[2] B. Barshan and H.F. Durrant-Whyte. Inertial navigation systems for mobile robots. *IEEE Transactions on Robotics and Automation*, 11(3):328–342, 1995.

[3] Kristijan Bartol, David Bojanić, Tomislav Petković, and Tomislav Pribanić. Generalizable human pose triangulation. In *Proceedings of IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 561–578. Springer International Publishing, Oct. 2016.

[5] Mazen Borno, Johanna O'Day, Vanessa Ibarra, James Dunne, Ajay Seth, Ayman Habib, Carmichael Ong, Jennifer Hick, Scott Uhlrich, and Scott Delp. Opensense: An open-source toolbox for inertial-measurement-unit-based measurement of lower extremity kinematics over long durations. *Journal of NeuroEngineering and Rehabilitation*, 19(22), 2022.

[6] Melissa Boswell, Łukasz Kidziński, Jennifer Hicks, Scott Uhlrich, Antoine Falisse, and Scott Delp. Smartphone videos of the sit-to-stand test predict osteoarthritis and health outcomes in a nationwide study. *medRxiv*, 2022.

[7] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[8] Mark Euston, Paul Coote, Robert Mahony, Jonghyuk Kim, and Tarek Hamel. A complementary filter for attitude estimation of a fixed-wing uav. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 340–345, 2008.

[9] Nima Ghorbani and Michael J. Black. SOMA: Solving optical marker-based mocap automatically. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11117–11126, Oct. 2021.

[10] Eni Halilaj, Soyong Shin, Eric Rapp, and Donglai Xiang. American society of biomechanics early career achievement award 2020: Toward portable and modular biomechanics labs: How video and imu fusion will change gait analysis. *Journal of Biomechanics*, 129:110650, 2021.

[11] Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415, 2016.

[12] Md Sanzid Bin Hossain, Joseph Dranetz, Hwan Choi, and Zhishan Guo. Deepbbwae-net: A cnn-rnn based deep superlearner for estimating lower extremity sagittal plane joint kinematics using shoe-mounted imu sensors in daily living. *IEEE Journal of Biomedical and Health Informatics*, pages 1–1, 2022.

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR.

[14] Marco Iosa, Pietro Picerno, Stefano Paolucci, and Giovanni Morone. Wearable inertial sensors for human movement analysis. *Expert Review of Medical Devices*, 13:641–659, 2016.

[15] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *International Conference on Computer Vision (ICCV)*, 2019.

[16] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):190–204, 2019.

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

[18] Robert M. Kanko, Elise K. Laende, Elysia M. Davis, W. Scott Selbie, and Kevin J. Deluzio. Concurrent assessment of gait kinematics using marker-based and markerless motion capture. *Journal of Biomechanics*, 127:110665, 2021.

[19] Meejin Kim and Sukwon Lee. Fusion poser: 3d human pose estimation using sparse imus and head trackers in real time. *Sensors*, 22(13), 2022.

[20] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015.

[21] Matthew M. Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 33(6):220:1–220:13, Nov. 2014.

[22] Sebastian O. H. Madgwick, Andrew J. L. Harrison, and Ravi Vaidyanathan. Estimation of imu and marg orientation using a gradient descent algorithm. In *2011 IEEE International Conference on Rehabilitation Robotics*, pages 1–7, 2011.

[23] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, Oct. 2019.

[24] Marion Mundt, William R. Johnson, Wolfgang Potthast, Bernd Markert, Ajmal Mian, and Jacqueline Alderson. A comparison of three neural network approaches for estimating joint angles and moments from inertial measurement units. *Sensors*, 21(13), 2021.

[25] Eduardo Palermo, Stefano Rossi, Francesca Marini, Fabrizio Patanè, and Cappa Paolo. Experimental evaluation of accuracy and repeatability of a novel body-to-sensor calibration procedure for inertial sensor-based gait analysis. *Measurement*, 52:145–155, 2014.

[26] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[27] Owen Pearl, Soyong Shin, Ashwin Godura, Sarah Bergbreiter, and Eni Halilaj. Fusion of video and inertial sensing data via dynamic optimization of a biomechanical model. *Journal of Biomechanics*, Under Review.

[28] Pietro Picerno. 25 years of lower limb joint kinematics by using inertial and magnetic sensors: A review of methodological approaches. *Gait Posture*, 51:239–246, 2017.

[29] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[30] Eric Rapp, Soyong Shin, Wolf Thomsen, Reed Ferber, and Eni Halilaj. Estimation of kinematics from inertial measurement units using a combined deep learning and optimization framework. *Journal of Biomechanics*, 116:110229, 2021.

[31] Iasonas Kokkinos Riza Alp Güler, Natalia Neverova. Densepose: Dense human pose estimation in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[32] Xavier Robert-Lachaine, Hakim Mecheri, Christian Larue, and André Plamondon. Accuracy and repeatability of single-pose calibration of inertial measurement units for whole-body motion analysis. *Gait Posture*, 54:80–86, 2017.

[33] Daniel Roetenberg, Henk Luinge, and Per Slycke. Xsens mvn: Full 6dof human motion tracking using miniature inertial sensors. 2009.

[34] Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1216–1224, 2017.

[35] Nils Roth, Martin Ullrich, Arne Küderle, Till Gladow, Franz Marxreiter, Heiko Gassner, Felix Kluge, Jochen Klucken, and Bjoern M. Eskofier. Real-world stair ambulation characteristics differ between prospective fallers and non-fallers in parkinson's disease. *IEEE Journal of Biomedical and Health Informatics*, 26(9):4733–4742, 2022.

[36] Nidhi Seethapathi, Shaofei Wang, Rachit Saluja, Gunnar Blohm, and Konrad P. Körding. Movement science needs different pose tracking algorithms. *CoRR*, abs/1907.10226, 2019.

[37] Kelly R. Sheerin, Duncan Reid, Denise Taylor, and Thor F. Besier. The effectiveness of real-time haptic feedback gait retraining for reducing resultant tibial acceleration with runners. *Physical Therapy in Sport*, 43:173–180, 2020.

[38] Patrick Slade, Ayman Habib, Jennifer L. Hicks, and Scott L. Delp. An open-source and wearable system for measuring 3d human motion in real-time. *IEEE Transactions on Biomedical Engineering*, 69(2):678–688, 2022.

[39] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[40] Matthew Trumble, Andrew Gilbert, Adrian Hilton, and John Collomosse. Deep autoencoder for combined human pose estimation and body model upscaling. In *European Conference on Computer Vision (ECCV'18)*, 2018.

[41] Matt Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *2017 British Machine Vision Conference (BMVC)*, 2017.

[42] Scott D. Uhlrich, Antoine Falisse, Łukasz Kidziński, Julie Muccini, Michael Ko, Akshay S. Chaudhari, Jennifer L. Hicks, and Scott L. Delp. Opencap: 3d human movement dynamics from smartphone videos. *bioRxiv*, 2022.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[44] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[45] Ge Wu, Sorin Siegler, Paul Allard, Chris Kirtley, Alberto Leardini, Dieter Rosenbaum, Mike Whittle, Darryl D. D'Lima, Luca Cristofolini, Hartmut Witte, Oskar Schmid, and Ian Stokes. Isb recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion - part i: Ankle, hip, and spine. *Journal of Biomechanics*, 35(4):543–548, 2002.

[46] Ge Wu, Frans C.T. van der Helm, H.E.J. (DirkJan) Veeger, Mohsen Makhsous, Peter Van Roy, Carolyn Anglin, Jochem Nagels, Andrew R. Karduna, Kevin McQuade, Xuguang Wang, Frederick W. Werner, and Bryan Buchholz. Isb recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion—part ii: shoulder, elbow, wrist and hand. *Journal of Biomechanics*, 38(5):981–992, 2005.

[47] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *CoRR*, abs/1505.00853, 2015.

[48] Zhe Zhang, Chunyu Wang, Wenhu Qin, and Wenjun Zeng. Fusing wearable imus with multi-view images for human pose estimation: A geometric approach. In *CVPR*, 2020.

[49] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.

[50] Lu Zhou, Yingying Chen, Yunze Gao, Jinqiao Wang, and Hanqing Lu. Occlusion-aware siamese network for human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 396–412, 2020.