

Tracking and Feedback in Surfing: A 3D Pose Estimation Approach

Master's Thesis in Computer Science

submitted
by

Ilias Masmoudi

born 29.08.1998 in Rabat

Written at

Machine Learning and Data Analytics Lab
Department Artificial Intelligence in Biomedical Engineering
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

Advisors: Alexander Weiß, M. Sc., Prof. Dr. Anne Koelewijn

Started: 20.09.2023

Finished: 20.03.2024

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Die Richtlinien des Lehrstuhls für Bachelor- und Masterarbeiten habe ich gelesen und anerkannt, insbesondere die Regelung des Nutzungsrechts.

Erlangen, den 20. März 2024

Abstract

This thesis presents a comprehensive study on the application of 2D and 3D pose estimation techniques to the dynamic and challenging environment of surfing, exploring the integration of inertial measurement unit (IMU) data to enhance model accuracy. Utilizing advanced computational models, this research navigates through the intricacies of pose estimation under conditions characterized by frequent occlusions and environmental variability inherent to water sports. Through a meticulous quantitative analysis, the study reveals that 2D pose estimation models, specifically ViTPose and RTMO, achieved confidence levels ranging between 80-90% across various camera views. The 3D pose estimation efforts were underscored by the finetuning of the MotionBERT model, which demonstrated a significant reduction in reprojection error, from an average of 1.51% before finetuning to 0.98% post-finetuning across all camera views. This improvement highlights the model's enhanced precision in capturing complex movements. However, the integration of IMU data did not yield the expected reduction in reprojection errors, suggesting that the added motion context from IMUs does not directly translate to improved accuracy in 3D pose estimation for surfing. This finding stimulates further discourse on the effectiveness of multimodal data integration in complex environments. Additionally, a comparative analysis revealed that the finetuned MotionBERT model outperformed a multi-view learnable triangulation model in reprojection error metrics, emphasizing the importance of temporal data in enhancing pose estimation accuracy. This thesis not only advances the technical understanding and capabilities of pose estimation models in sports and dynamic activities but also lays the groundwork for future research in multimodal data integration and the development of more sophisticated pose estimation techniques. The findings hold profound implications for athletic training, performance enhancement, and injury prevention, promising to contribute significantly to sports science and beyond.

Contents

1	Introduction	1
2	Literature Review	5
2.1	2D Pose Estimation	6
2.2	3D Pose Estimation	11
2.2.1	Monocular 3D Pose Estimation	11
2.2.2	Multi-View 3D Pose Estimation	14
2.2.3	IMU-Assisted 3D Pose Estimation	17
2.3	Evaluation Metrics	20
2.3.1	2D Human Pose Estimation	20
2.3.2	3D Human Pose Estimation	22
3	Methods	25
3.1	Data Collection	26
3.2	Data Processing	30
3.2.1	Multi-View Camera Synchronization	31
3.2.2	Multi-View Camera Calibration	31
3.2.3	IMU-Video Synchronization	32
3.3	2D Pose Estimation	33
3.4	3D Pose Estimation	35
3.4.1	Single-View 3D Pose Estimation	35
3.4.2	Exploration of Single-View 3D Pose Prediction Enhancement Using The Surf Pose Data	36
3.4.3	Exploration of Single-View 3D Pose Prediction Enhancement Through IMU Integration	37
3.4.4	Multi-view 3D Pose Estimation	39

4	Results	45
4.1	2D Pose Estimation Results	45
4.2	3D Pose Estimation Results	47
4.2.1	Evaluating IMU Data Integration on Reprojection Error	50
4.2.2	Comparison with Multi-view 3D Pose Estimation	51
5	Discussion	59
6	Conclusion	61
	List of Figures	63
	List of Tables	65
	Bibliography	67
A	Acronyms	73

Chapter 1

Introduction

Surfing, a sport with ancient Polynesian origins, notably in Hawaii, has transformed into a global sensation, deeply rooted in diverse cultures and communities across the world. This sport transcends mere leisure activity, embodying profound spiritual and cultural significances that have been cherished for centuries. An illustration from the late 18th century, found in Captain Cook's Voyages, provides a glimpse into the early days of Hawaiian surfing, hinting at a rich history that predates the sport's temporary decline in the 19th century and its subsequent resurgence (see Fig. 1.1)[Nen17; Coo97]. The global surfing market, projected to reach nearly five billion U.S. dollars by 2027, highlights the sport's vast professional appeal and widespread recreational participation. Further signifying its growing stature, surfing's inclusion in the Tokyo 2020 Olympic Games marked a milestone, elevating the sport's professional image and global presence. Additionally, its expected appearance in the 2024 Paris Olympics and the possibility of being featured in the 2028 Paralympic Games underscore its broadening appeal and recognition on the international stage [Dep24].

Technological advancements have not only reshaped the geographical landscape of surfing but have also led to innovations in equipment and training methods. The advent of river surfing, leveraging the natural flow of rivers to create standing waves, exemplifies the sport's expansion into non-coastal areas. This specialization necessitates training techniques that address the unique conditions of river surfing environments, setting a foundation that could eventually be adapted to ocean surfing contexts [ZöL23]. This shift underscores the necessity for training methods that integrate technologies to enhance performance and injury prevention across both river and potentially ocean surfing scenarios [Bad21].

Technological advancements have not only reshaped the geographical landscape of surfing but have also led to innovations in equipment and training methods. The advent of river surfing,

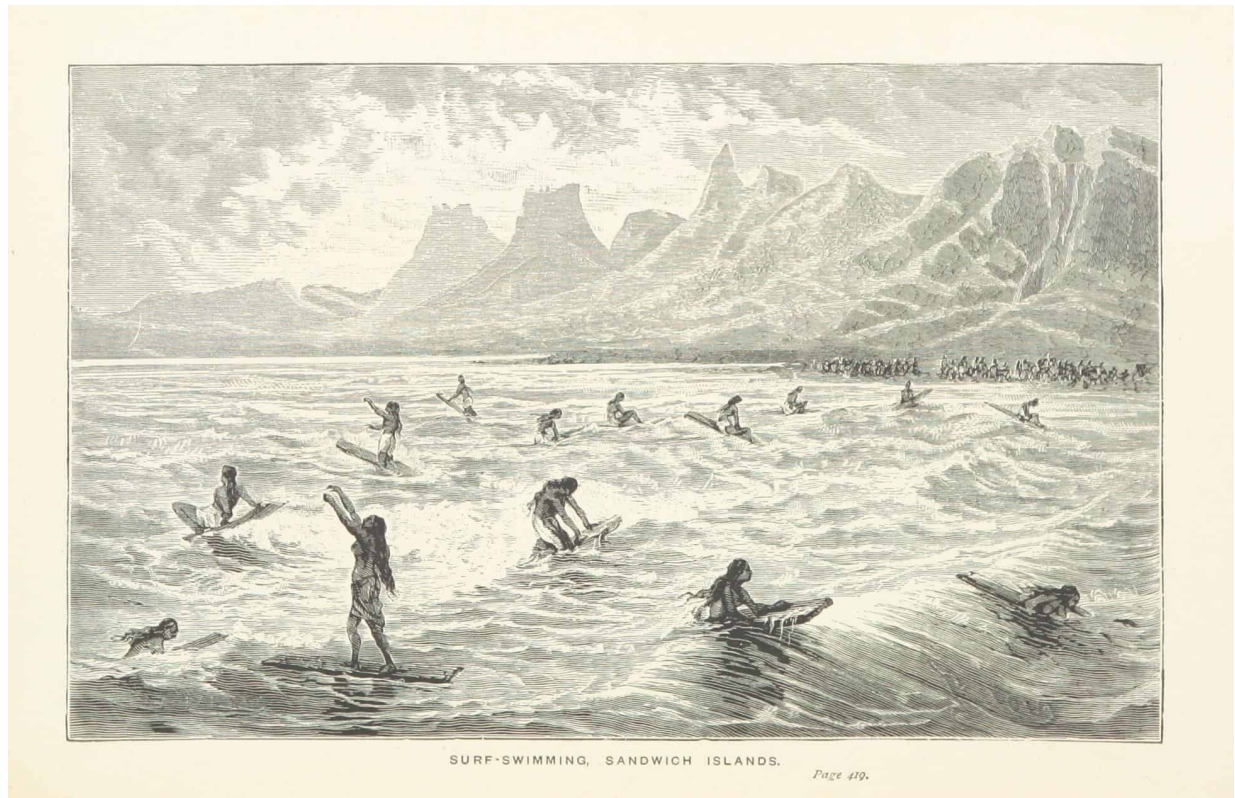


Figure 1.1: 1897 depiction from Captain Cook's Voyages, illustrating early Hawaiian surfing, prior to the sport's 19th-century suppression and later resurgence. [Coo97]

leveraging the natural flow of rivers to create standing waves, exemplifies the sport's expansion into non-coastal areas. This specialization necessitates training techniques that address the unique conditions of river surfing environments, setting a foundation that could eventually be adapted to ocean surfing contexts [ZÖL23]. The adaptation of these training techniques to ocean surfing is facilitated by understanding the fundamental differences between the two environments: ocean waves present a changing environment, and surfers need to paddle to catch waves, both aspects that are less prevalent in river surfing. Addressing these differences is crucial for developing comprehensive training methods that can enhance performance and injury prevention across both river and potentially ocean surfing scenarios [Bad21; Bor18].

Traditional surfing instruction, characterized by subjective analysis and personalized coaching, often falls short in providing consistent and objective feedback, which in this context means utilizing quantifiable measures such as kinematics (movement patterns), kinetics (forces involved), acting forces on the surfer and board, and muscle activations. This gap is especially pronounced in river surfing, where the dynamic and unpredictable nature of the waves challenges the efficacy

of conventional coaching methods. The variability of these conditions demands a more sophisticated approach to performance feedback [Bad21].

Addressing this need, this thesis proposes the development of an advanced pose estimation system initially focusing on river wave surfing but designed with the flexibility to be applied across a variety of sports, especially those occurring within defined volumes. While acknowledging the utility of IMUs in providing detailed motion data, this research recognizes their limitation due to drift over time, which can lead to inaccuracies in long-duration monitoring of athletic performance. The proposed pose estimation system aims to bridge the gap in objective performance feedback, creating a tool that, while adaptable to the surfing context, offers a non-invasive and potentially accurate means of assessing athlete movement. This system also holds promise for broader applications in diverse sporting disciplines, suggesting the potential for future adaptations to sports beyond surfing, where dynamic movements and precise performance feedback are crucial. The objectives include designing, implementing, and validating the pose estimation system, thereby setting the stage for future advancements in sports technology that could benefit a wide range of physical activities [Far17].

The motivation behind this research is driven by the potential benefits of an accurate pose estimation system in enhancing training quality, safety, and performance in surfing across different environments. By providing surfers with detailed, objective feedback on their technique, the system can facilitate precise adjustments, leading to more efficient training, reduced injury risk, and improved performance outcomes in river settings [Wen23]. Moreover, this research is positioned to contribute significantly to the field of sports technology by exploring the application of computer vision and machine learning in analyzing and enhancing athletic performance in water sports.

The development of a sophisticated, markerless pose estimation system for surfing represents a promising avenue for advancing the sport's technical training and safety protocols. By addressing the unique challenges of motion analysis in aquatic environments, this research has the potential to make a substantial impact on the field of sports technology, offering insights and innovations that could benefit athletes across a range of water sports.

Chapter 2

Literature Review

At the forefront of advancing computer vision technologies, human pose estimation (HPE) emerges as a paramount field, especially with the integration of deep learning. This area of research is pivotal for developing systems capable of understanding and interpreting human body configurations in two and three dimensions, laying the groundwork for numerous applications, from augmented reality to enhanced athletic training. The evolution of HPE methodologies, driven by deep learning innovations, has opened new avenues for analyzing complex human activities [Bri19] [Zhe20], such as river wave surfing. This activity presents unique challenges due to its dynamic nature and the fluid interaction between the surfer and the fluid environment. This comprehensive examination not only charts the progress in the field but also sets the stage for exploring the specific requirements and potential solutions for accurately estimating poses in the context of river wave surfing, illustrating the intricate balance between general HPE principles and their application to specialized scenarios.

In the realm of two-dimensional (2D) HPE, methods are primarily split according to the target number of individuals: either focusing on a single person or accommodating multiple persons within the scene. For the single-person scenario, two primary methodologies emerge: regression techniques, which directly map input features to pose coordinates in the image frame, and heatmap-based strategies, where the pose is inferred from heatmaps that represent the likelihood of joint locations.

Extending HPE to scenarios involving multiple individuals, 2D methodologies are sub-divided into top-down and bottom-up approaches. Transition to three-dimensional (3D) HPE, is distinguished by the nature of its input sources. Methods utilize either monocular RGB images and videos or alternative sensing technologies, such as inertial measurement units [Von18; Hua20a]. The predominant focus within 3D HPE is on monocular inputs, given their wide availability and

the rich contextual information they offer [Zhe20]. These approaches are subdivided based on the viewing configuration and the number of subjects present: single-view single-person, single-view multi-person, and multi-view methods. Notably, multi-view setups are especially beneficial in multi-person HPE, leveraging multiple angles to enhance accuracy and overcome occlusions, although specific distinctions between single and multi-person configurations are not made within this category [Isk19].

This structured overview of HPE methodologies underscores the diverse strategies employed to capture human poses across various dimensions and settings, highlighting the field's complexity and the nuanced considerations required when designing and implementing deep learning-based pose estimation systems.

2.1 2D Pose Estimation

The advent of deep learning has considerably transformed the landscape of 2D HPE, marking a pivotal shift from conventional methodologies that relied on manual feature extraction and simplistic representations of the human form. This evolution underscores the transition towards models that are capable of understanding the intricate geometries and dynamics of human movement through images and videos. The focus is primarily on identifying and localizing key points that delineate human body joints, thereby enabling the reconstruction of the human pose in a two-dimensional plane [Bur13].

There are two main pose estimation approaches: top-down or bottom-up. The principal issue of the latter method is that recovery from a failed person detection is difficult. Bottom-up approaches on the other hand, attempt to reduce the run-time complexity from the number of people by directly detecting all the parts from the images and then, associating them with each individual. Fig. 2.1 illustrates the differences between the two paradigms.

Pose estimation algorithms have considerably improved throughout the last few years. One very established model is OpenPose [Cao17]. OpenPose represents a groundbreaking approach in the field of computer vision for the detection and analysis of human poses from 2D images. The code of its methodology is a sophisticated two-branch multi-stage convolutional neural network (CNN) architecture designed to accurately predict the 2D locations of anatomical keypoints on multiple people within a single image (see Fig. 2.2). The system inputs a color image and outputs the 2D pixel coordinates of body keypoints for each person detected. The first branch of the network generates confidence maps for body part locations, while the second branch predicts part affinity fields (PAFs), a novel representation that encodes the degree of association between

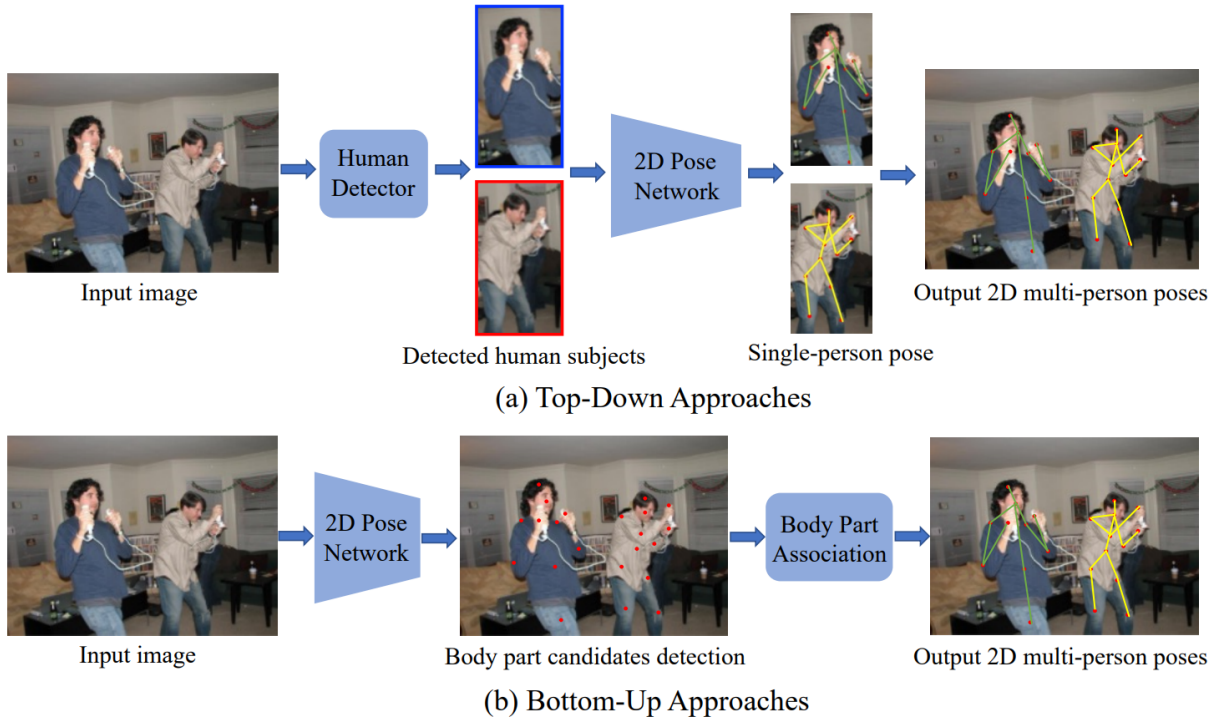


Figure 2.1: Diagram displaying frameworks for 2D HPE involving multiple persons. (a) The top-down methodology includes two distinct phases: firstly, identifying individual humans within the scene, and secondly, estimating the pose for each detected human region; (b) The bottom-up strategy involves initially detecting potential keypoints for all body parts across the scene and subsequently linking these parts across different figures to construct unique poses for each person, as depicted by [Zhe20].

parts to accurately capture the spatial relationship between limbs. These PAFs are crucial for differentiating between overlapping or closely situated individuals by preserving both the position and orientation of each limb across the image. This dual output is then processed through a series of stages, refining the predictions by integrating the output of both branches at each stage with the original image features. This iterative refinement process, coupled with intermediate supervision to counteract the vanishing gradient problem, enables OpenPose to achieve remarkable precision in real-time pose estimation. Furthermore, this method introduces a novel approach to part association, leveraging the geometric and orientation information encoded in PAFs to assemble detected body parts into coherent full-body poses for an unknown number of people, addressing the challenges posed by crowded scenes. The efficacy of OpenPose, achieving a 56.3% average precision (AP) on the COCO-WholeBody dataset, is underscored by its ability to perform multi-

person parsing and part association with high accuracy, making it a pivotal contribution to the advancement of pose estimation technologies [Cao17]. However, while OpenPose has indeed made a pivotal contribution to the advancement of pose estimation technologies, it is not without its limitations. The model can exhibit failure cases, particularly when dealing with complex poses, occlusions, or low-resolution images. In such scenarios, the accuracy of part detection and association can be compromised, leading to erroneous pose estimation. This highlights the continuing need for research in this field to address these challenges and improve upon the existing technology.

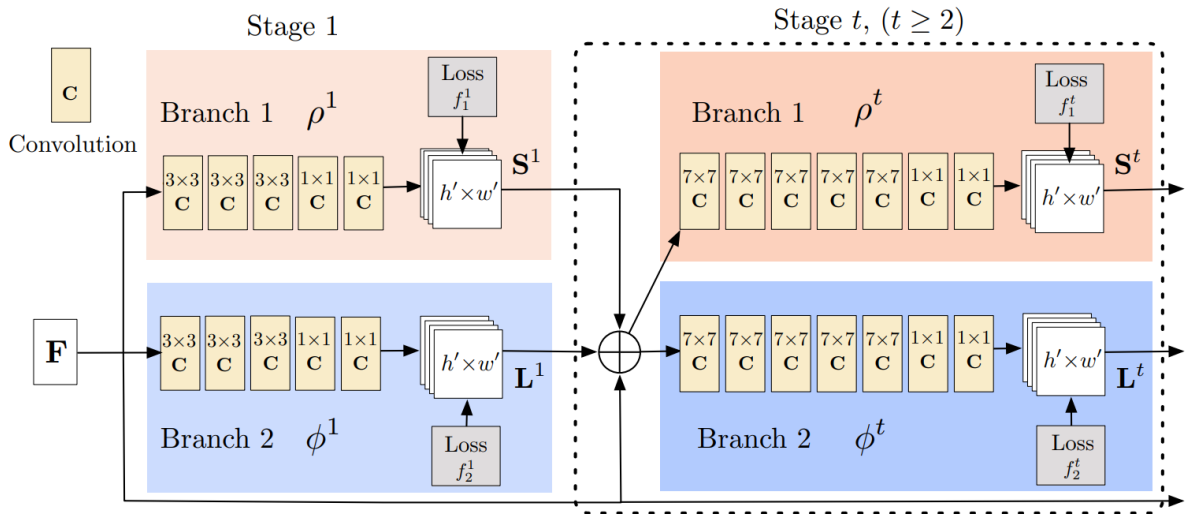


Figure 2.2: Overview of the Multi-Stage CNN for Pose Estimation: This architecture illustrates the iterative process of predicting confidence maps and PAF across multiple stages. Each stage refines the accuracy of both confidence map (S_t) predictions for keypoint localization and PAF (L_t) predictions for limb connectivity, leveraging the combined insights from previous stages and image features for continuous enhancement, as depicted by [Cao17].

In their study on 2D pose estimation, Jiang et al. [Jia23] introduce RTMPose, a model that notably advances the field by incorporating efficient backbone architectures within the coordinate classification paradigm. The concept of coordinate classification, as enhanced from SimCC [Li22], innovatively transforms keypoint localization into a classification problem. This is achieved by discretizing the continuous space of keypoint locations into a finite set of categories, effectively converting the prediction of precise coordinate values into a classification task. The approach significantly benefits model performance through strategic adjustments in training procedures and the integration of advanced modules and micro-designs. Notably, RTMPose, achieving a 71.4% AP on the COCO-WholeBody dataset [24a], employs a Gaussian label smoothing technique, in-

spired by traditional classification tasks, to considerably improve accuracy. By adopting a more compact backbone and optimizing the training regimen with techniques such as pre-training, exponential moving average, and two-stage training augmentations [Hua20b; Lyu22], the model achieves remarkable precision. As illustrated in Fig. 2.3, further enhancements include the use of self-attention modules to refine keypoint representations and strategic modifications to the loss function, addressing coordinate classification as an ordinal regression task. The study also explores the efficacy of different convolutional kernel sizes and the impact of training duration and dataset diversity on model performance. Ultimately, RTMPose culminates in an optimized inference pipeline, incorporating mechanisms for lower latency and enhanced robustness in real-world applications. This detailed approach to enhancing 2D pose estimation models represents a notable progression in accuracy and efficiency, particularly within the arduous context of surfing, and offers important directions for future research in the area.

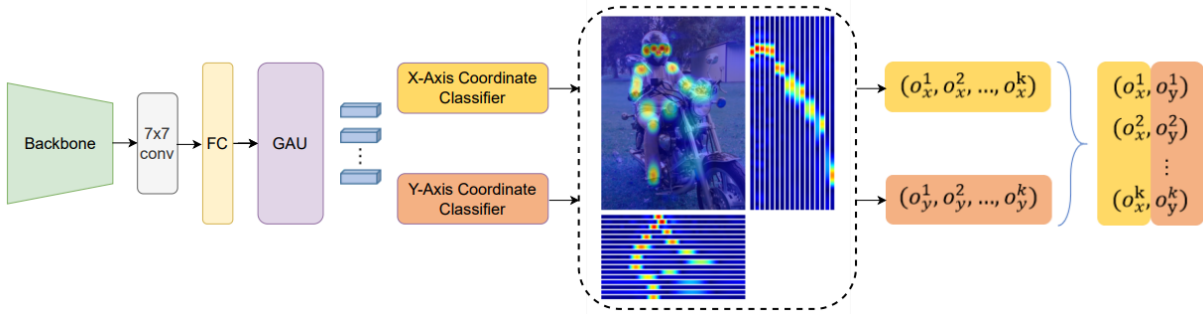


Figure 2.3: The RTMPose Architecture Overview. [Jia23] highlighting its core components: a convolutional layer for initial feature extraction, a fully-connected layer for processing those features, and a GAU dedicated to refining representations of K keypoints. The architecture innovatively approaches 2D pose estimation by treating the determination of x-axis and y-axis coordinates as separate classification tasks, thereby facilitating precise prediction of both horizontal and vertical keypoint locations.

The advancements in 2D pose estimation are further enriched by the introduction of DW-Pose [Yan23], which represents a novel approach through its two-stage pose distillation (TPD) strategy. Building upon the foundation set by RTMPose, DWPose integrates a first-stage distillation where a pre-trained teacher model guides the student model from its inception. This guidance occurs at both the feature and logit levels. This process involves meticulously aligning the student’s feature maps and logits with those of the teacher’s, using mean squared error (MSE) loss and a logit-based distillation method to ensure that the student accurately mimics the teacher’s outputs. The second stage introduces a self-knowledge distillation (KD) approach where the model

employs its own logits to refine the training of its head, bypassing the need for labeled data. This innovative methodology considerably accelerates training efficiency and boosts performance, especially in keypoint localization tasks. Through the employment of a weight-decay strategy, the DWPose method balances distillation and original loss components to optimize learning. The second-stage distillation further capitalizes on the trained student model to self-improve, highlighting a shift towards more autonomous learning frameworks in pose estimation technologies. This paper’s contribution, especially its head-aware distillation technique, marks a considerable leap towards reducing training time while enhancing the precision of pose estimation models. Despite its notable contributions, DWPose faces certain limitations that merit discussion. One limitation of DWPose is its dependency on a high-quality teacher model for the initial distillation phase. This reliance could constrain its applicability in scenarios where such a model is not readily available or where the domain-specific nuances of the dataset might not be fully captured by the teacher model. Additionally, while the method shows significant improvements on datasets like COCO-WholeBody and UBody [Jin20; Lin23], the generalizability of these gains to a broader array of datasets and real-world scenarios remains to be extensively evaluated.

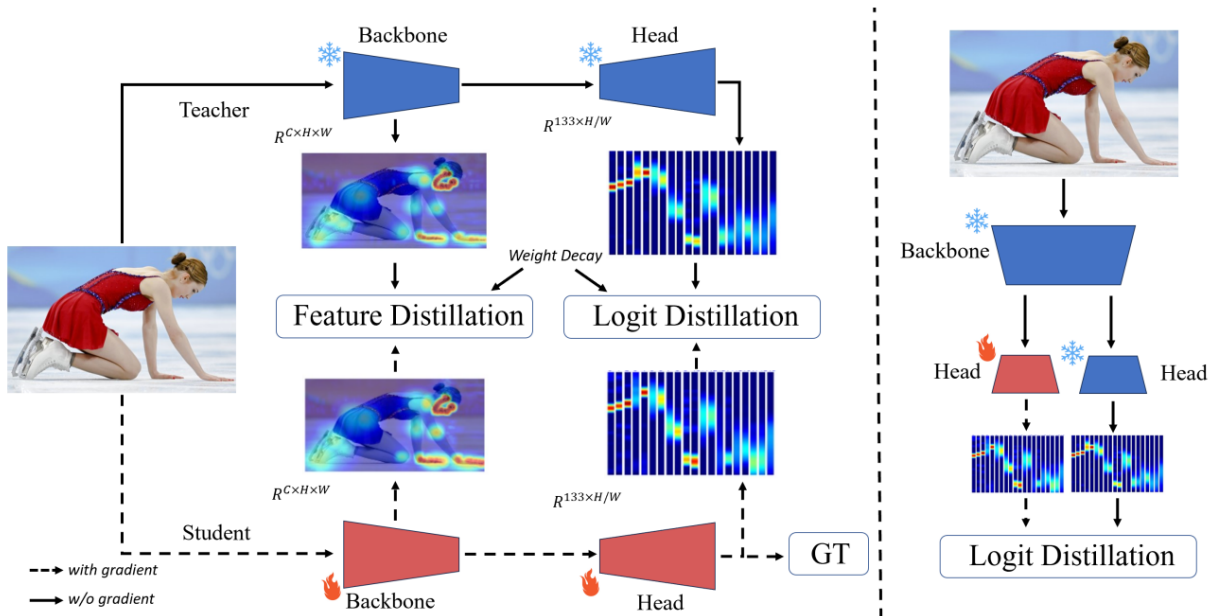


Figure 2.4: TPD Pipeline of DWPose. On the left, the initial stage of distillation utilizes a conventional approach that incorporates both feature and logit levels. On the right, the subsequent stage of distillation utilizes the student model to instruct a new head. [Yan23]

2.2 3D Pose Estimation

In recent years, the study of 3D HPE has emerged as a pivotal area of research due to its capacity to map human body joints in three dimensions. This advancement holds promise for a variety of fields, including virtual reality, animation, and sports science, [Wen19; Wan19] offering a deeper understanding of human movement. Despite strides made in 2D HPE, the leap to 3D HPE poses unique challenges, primarily because it relies on data that reduces three-dimensional reality to two-dimensional images or videos. This simplification inherently omits crucial depth information, complicating the accurate depiction of human poses in three dimensions [Zhe20].

Addressing these challenges often involves leveraging additional data sources, such as multiple camera angles or sensors like IMUs, to regain lost dimensional information through techniques of sensor fusion [Von18; Hua20a]. Yet, the dependency on extensive, and often environment-specific, data sets for training deep learning models introduces further obstacles. The manual collection and annotation of 3D data are notably labor-intensive, with a tendency towards indoor settings and limited actions, which restricts the versatility and applicability of trained models and thus hard to infer in real-world-scenarios, especially when complex circumstances or environments are given.

This section aims to explore the methodologies and innovations in 3D HPE, firstly emphasizing on single-view approaches, followed by multi-view approaches. Eventually, expanding to consider methods that integrate additional sensory inputs.

2.2.1 Monocular 3D Pose Estimation

In the rapidly evolving field of 2D to 3D pose uplifting, the diffusion-based 3D pose estimation (D3DP) method represents a significant advancement, introducing a novel approach that leverages diffusion models for generating multiple high-quality 3D pose hypotheses from single 2D observations [Sha23]. As depicted in Fig. 2.5, the core of D3DP involves a two-step Markov chain process: an initial diffusion phase that introduces Gaussian noise to degrade the ground truth 3D poses, followed by a reverse process where a denoiser, conditioned on 2D keypoints and specific time steps, reconstructs the noise-free 3D poses. This technique not only facilitates the generation of diverse pose hypotheses but also allows for the customization of the number and quality of these hypotheses through adjustable parameters, addressing the limitations of non-adjustable hypothesis counts found in previous methodologies.

Moreover, the introduction of the joint-wise reprojection-based multi-hypothesis aggregation (JPMA) method marks a considerable stride towards practical application. Using JPMA

one can select the best hypothesis for each joint based on its proximity to the corresponding 2D keypoint, allowing for a more accurate and reliable 3D pose prediction by combining the strengths of individual hypotheses at the joint level. This approach contrasts with traditional methods that either focus on pose-level selection, often leading to suboptimal joint accuracy, or do not fully exploit the geometric information contained in 2D keypoints.

D3DP’s architecture is designed to be compatible with current deterministic 3D pose estimators [Sha23], using them as the backbone for the denoiser. This compatibility is achieved through minor modifications, such as the integration of 2D keypoints as additional guidance for the denoising process and the incorporation of timestep embedding to inform the denoiser of the noise level. These enhancements enable D3DP to maintain computational efficiency while significantly improving the flexibility and accuracy of 3D pose estimation from 2D data, making it a noteworthy contribution to the literature on pose estimation technologies.

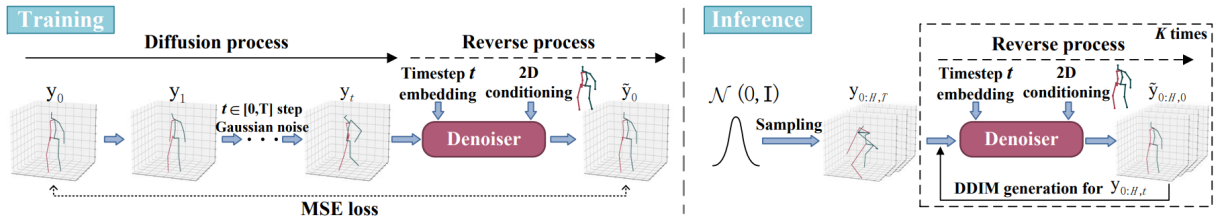


Figure 2.5: Overview of the proposed D3DP method. Left: Training phase, where t -step Gaussian noise is added to the ground truth 3D pose, resulting in the noisy pose. This pose is then processed by a denoiser conditioned on 2D keypoints x and timestep t to produce the final prediction. Right: Inference phase, starting with H samples drawn from a Gaussian distribution to initialize 3D poses $y_{0:H,T}$, which are refined to generate the noiseless 3D pose hypotheses. Additionally, the reverse process can be iterated K times to refine the final results by processing DDIM-generated 3D poses with varying noise levels through the denoiser, as depicted from [Sha23]

In the evolving landscape of 3D monocular pose estimation, the MotionBERT [Zhu23] model emerges as a considerable innovation, delineating a path that might well redefine the standards of accuracy and efficiency in the field. At the heart of MotionBERT lies a meticulously designed two-stage methodology that encompasses both unified pretraining and task-specific finetuning, with the dual-stream spatio-temporal transformer (DSTformer) architecture playing a pivotal role in executing the critical 2D-to-3D lifting task. This approach is reflective of a deep understanding of the complexities involved in human motion capture and analysis.

The initial stage of MotionBERT’s methodology utilizes 2D skeleton sequences as input, a choice underpinned by the sequences’ proven reliability across various motion sources and their

robustness against environmental and situational variations. This strategic decision to use 2D skeletons paves the way for a more inclusive and versatile model, capable of handling data from a wide array of sources without compromising on the accuracy of the 3D pose estimations it generates.

Further integrating into its architecture, MotionBERT introduces a dual-stream Spatio-temporal Transformer that ingeniously captures both the intra-frame and inter-frame dynamics of human motion. The DSTformer, incorporating spatial and temporal blocks, adeptly models the complex interactions among body joints, thereby enhancing the model’s precision in 3D pose reconstruction. This unified pretraining framework addresses critical challenges in 3D pose estimation: developing a powerful motion representation that generalizes across tasks and efficiently utilizing heterogeneous human motion data. Inspired by successful practices in language and vision modeling, MotionBERT’s task of 2D-to-3D lifting not only facilitates the learning of comprehensive motion representations but also enables the effective incorporation of large-scale 3D motion capture (MoCap) data into the training process. The task-specific finetuning phase of MotionBERT’s methodology is characterized by its minimalist design principle. By implementing a shallow downstream network, the model maintains a focus on efficiency without sacrificing performance. This phase allows MotionBERT to apply its prelearned 3D-aware and temporal-aware human motion representation to a variety of downstream tasks, including 3D pose estimation, skeleton-based action recognition, and human mesh recovery. Each application showcases the model’s versatility and its potential to significantly improve the accuracy and reliability of pose estimations in real-world scenarios.

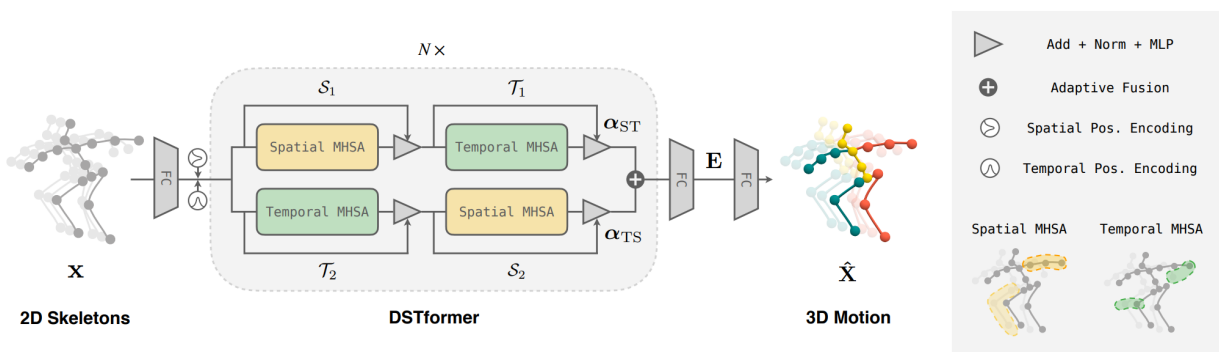


Figure 2.6: Architecture of the Model: DSTformer features N dual-stream-fusion modules, each incorporating pathways for spatial and temporal MHA mechanisms, complemented by MLPs. Spatial MHA captures inter-joint relationships at individual time steps, while temporal MHA tracks the movements of each joint across time.[Zhu23]

Building on the exploration of 3D monocular pose estimation models within this literature review, it's essential to introduce MotionAGFormer [Meh23], a noteworthy advancement that encapsulates the fusion of transformer and graph convolutional networks (GCNs). This innovative approach underlines a significant leap forward in the quest to lift 2D skeleton sequences to 3D pose sequences with heightened accuracy and efficiency. Central to the MotionAGFormer methodology is the concept of the MetaFormer, an evolved Transformer architecture that diverges from traditional attention mechanisms. By incorporating any token mixer, such as MHSA and GCNs, the MetaFormer framework offers a versatile foundation for processing and analyzing skeletal data. This flexibility and innovation in skeletal data analysis are visually depicted and further explained in the architecture figure (see Fig. 2.6)

The MotionAGFormer architecture (see Figure 2.7) is distinguished by its dual-stream spatio-temporal blocks, meticulously designed to capture intricate spatial relationships among individual body joints and the dynamic temporal relationships across sequential frames. This dual approach leverages a linear projection layer to transform 2D input sequences into a rich, high-dimensional feature space. Spatial position embeddings are then integrated, ensuring that the model retains critical positional information throughout the estimation process. The architecture employs AGFormer blocks, which are comprised of Spatial and Temporal MetaFormers within each stream, to further refine the understanding of the 3D structure embedded within skeletal sequences.

A key innovation within the MotionAGFormer model is its use of adaptive fusion techniques to synergize features from both Transformer and GCN streams. This method allows for an intelligent blend of spatial and temporal data, optimizing the model's ability to discern and predict 3D poses with remarkable precision. The effectiveness of this approach is further underscored by a sophisticated loss function that simultaneously addresses position accuracy and motion smoothness, ensuring that the model can accurately reflect both the static postures and the fluid movements inherent in human motion.

2.2.2 Multi-View 3D Pose Estimation

Transitioning from monocular to multi-view 3D pose estimation, the work by Iskakov et al.[Isk19] introduces an innovative methodology that leverages synchronized video streams from multiple cameras to estimate the global 3D positions of human joints. Unlike approaches that rely on single-view inputs, this method utilizes predefined projection matrices and processes images independently across frames without temporal information, focusing on a fixed set of human joints. Key to their approach is the cropping of images around the subject using either detected or ground-truth bounding boxes, followed by processing through a deep convolutional neural network architecture.

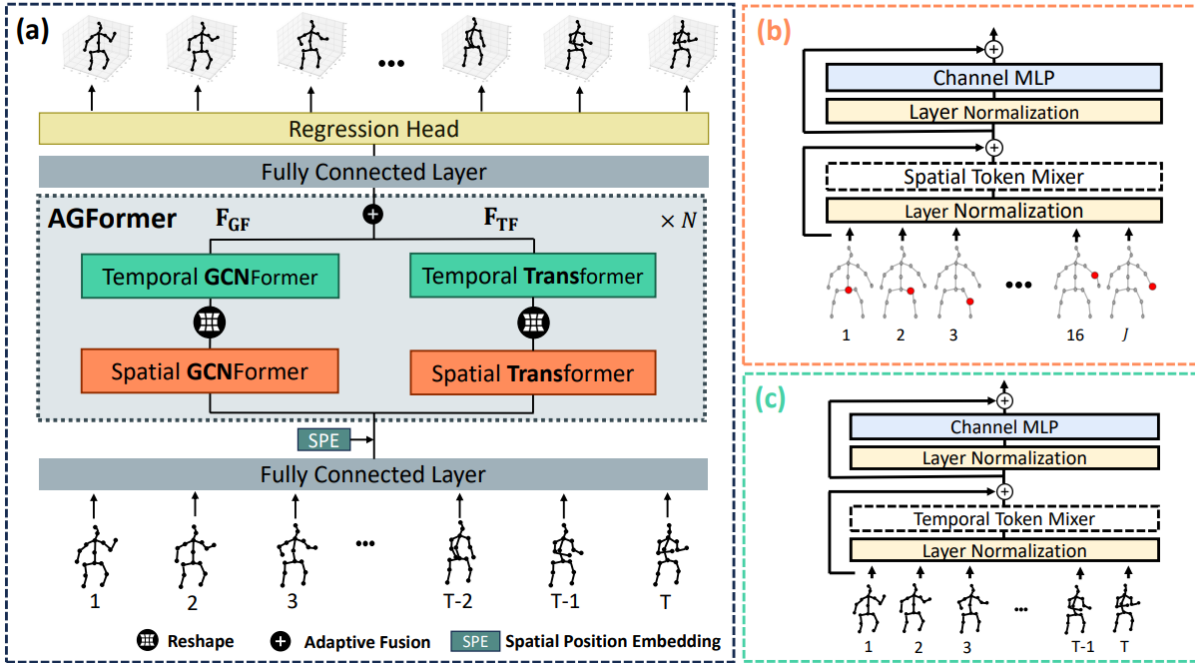


Figure 2.7: (a) Architecture of MotionAGFormer which is composed of N dual-stream spatio-temporal blocks. In this setup, one stream utilizes Transformers, while the other stream makes use of GCNFormers. (b) The Spatial MetaFormer, where every input token is associated with a specific joint in the human body. (c) The Temporal MetaFormer, with input tokens corresponding to individual frames within a sequence of poses, as depicted by [Meh23]

This network produces joint heatmaps from which 2D positions are derived using a soft-argmax operation, enabling gradient backpropagation and allowing for a more nuanced adjustment of heatmap responses early in training. This process is visually depicted in Fig. 2.8

The paper details two novel triangulation techniques for inferring 3D joint coordinates from these 2D estimations. The first, an algebraic triangulation method, predicts joint positions from different camera views, incorporating learnable weights to mitigate the impact of unreliable estimates due to occlusions or other factors. The second, a volumetric triangulation approach, addresses the baseline method’s limitations by projecting 2D backbone feature maps into 3D volumes, which are then processed to produce 3D joint heatmaps. This method not only facilitates the addition of a 3D human pose prior but also effectively filters out erroneous camera views, leading to more accurate pose estimations compared to the algebraic method.

Both methods enable end-to-end training, allowing for the backpropagation of gradients and enhancing training robustness with tailored loss functions. These innovations ensure heatmap

interpretability and, notably, the volumetric model achieved an average mean per joint position error (MPJPE) of 17.7 mm on the Human3.6M dataset, showcasing its precision in 3D human pose estimation.

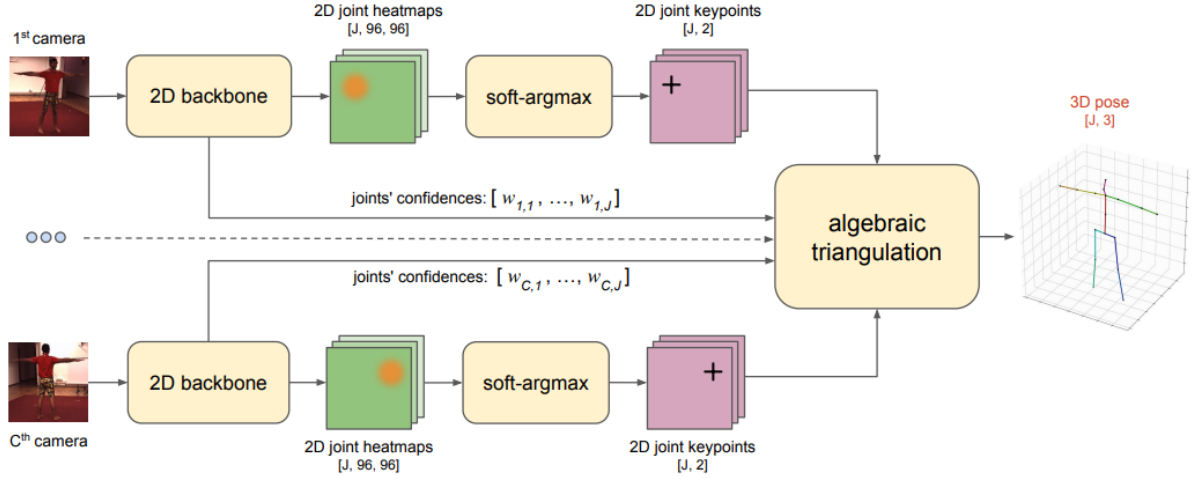


Figure 2.8: Diagram illustrating the method utilizing algebraic triangulation complemented by learned confidence levels for pose estimation. This technique starts with an array of RGB images, each accompanied by predefined camera parameters. A 2D neural network backbone processes these images to generate joint heatmaps and evaluates the confidence levels for each camera-joint combination. Using a soft-argmax operation, the model deduces the 2D coordinates of joints from these heatmaps. These coordinates, along with their associated confidence scores, are then input into an algebraic triangulation process to reconstruct the 3D pose. The architecture supports the backward propagation of errors, facilitating end-to-end training of the model, as depicted by [Isk19].

Building on the foundation of current 3D multi-view pose estimation models, the paper by Chun et al. [Chu23a] proposed a method which aims to reconstruct an skinned multi-person linear model (SMPL)-based [Lop15] 3D mesh of a single person from calibrated multi-view images through a complex yet efficient process. At the core of this approach is the innovative use of vertex heatmap autoencoders (VHAs) coupled with a body code predictor (BCP) and a fitting module. The VHA encoder transforms volumetric heatmaps into a low-dimensional latent code, which is then used to reconstruct the heatmaps and subsequently derive the 3D coordinates of subsampled mesh vertices via a 3D soft-argmax operation. This process is critical for maintaining the accuracy of the reconstructed mesh while managing computational efficiency.

Further delving into the specifics, the paper outlines a detailed structure for the VHA, incorporating a blend of basic convolution blocks, residual blocks, downsample blocks, and $1 \times 1 \times 1$

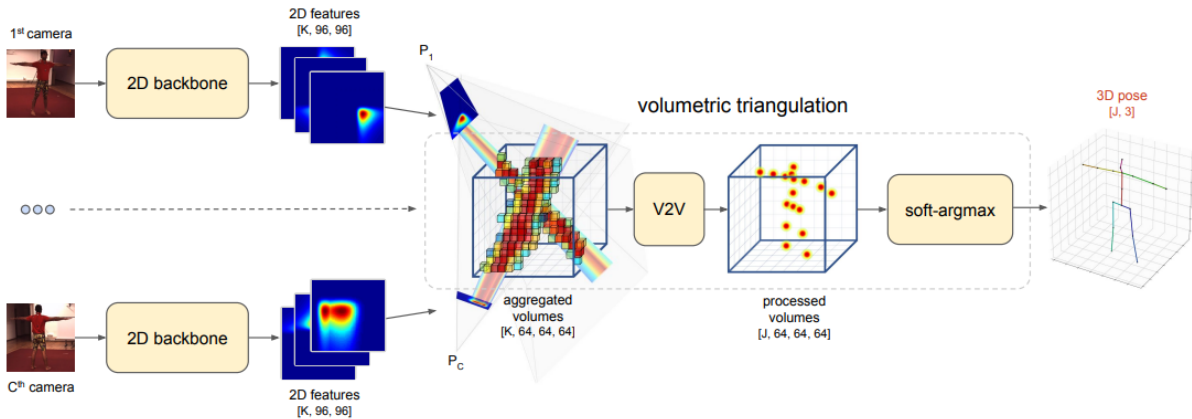


Figure 2.9: Conceptual framework of the volumetric triangulation method: This technique commences with a collection of RGB images, each annotated with precise camera parameters. The initial processing stage employs a 2D convolutional backbone to generate intermediary feature representations, which are then projected into 3D space, forming volumes. These volumes undergo an aggregation process, resulting in a consolidated volume of fixed dimensions. This processed volume is further analyzed by a 3D CNN, yielding 3D joint heatmaps. The final 3D joint positions are deduced through the application of a soft-argmax function on these heatmaps, as depicted by [Isk19].

convolution blocks, illustrating a methodical approach towards capturing the intricacies of human mesh reconstruction. Additionally, the BCP, borrowing elements from the learnable human mesh triangulation (LMT) model, as illustrated in Fig 2.10, replaces vertex regression with a 3D CNN encoder to predict the latent code from multi-view images, marking a significant shift towards more refined and precise mesh reconstruction techniques, as visually represented in Fig. 2.10. This transition from traditional vertex heatmap generation to latent code prediction exemplifies the paper’s contribution to enhancing the accuracy and fidelity of 3D human mesh models. Such advancements not only underscore the potential for more realistic and dynamic pose estimation in complex environments but also pave the way for future research in the domain of 3D human form and motion analysis. Notably, the model achieves an Average MPJPE of 15.6 mm on the Human3.6M dataset, further substantiating the model’s efficacy in accurately estimating human poses in three-dimensional space.

2.2.3 IMU-Assisted 3D Pose Estimation

In the evolving landscape of 3D human pose estimation, the IMUs with visual data marks an important advancement towards achieving high fidelity in capturing human motion in uncontrolled

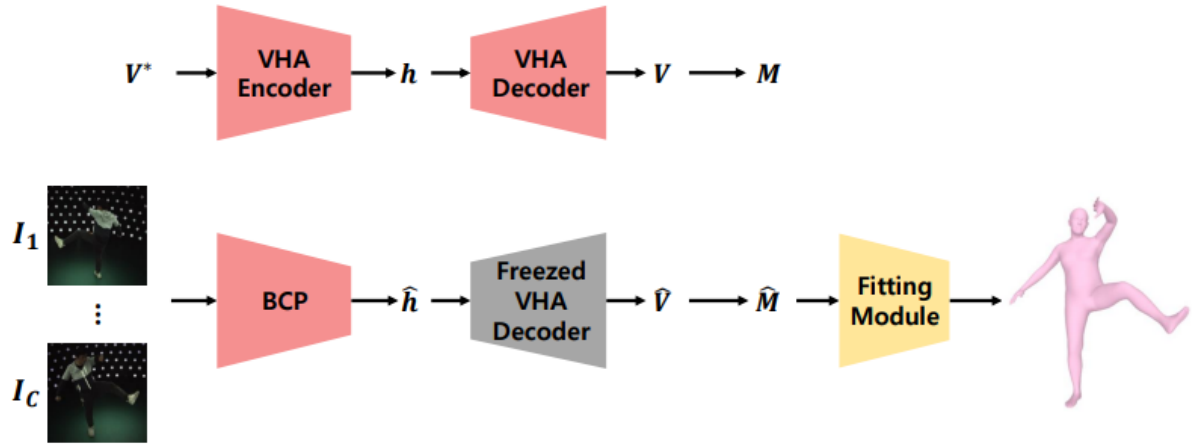


Figure 2.10: Schematic representation of the proposed 3D human mesh reconstruction method, featuring the VHA) that converts ground-truth subsampled vertices into a compact latent code and back to volumetric heatmaps for precise vertex localization. The BCP generates this latent code from multi-view images, which is then decoded to reconstruct the 3D mesh vertices. A fitting module subsequently adjusts the SMPL model parameters to the reconstructed mesh, ensuring accurate human pose representation, as depicted by [Chu23a]

environments. Building upon the foundation laid by monocular and multi-view pose estimation models, the work proposed by Marcard et al. [Von18] presents a novel methodology that addresses several limitations inherent to existing IMU-based MoCap systems. Traditional approaches, such as those employing Kalman Filters with multiple IMUs or custom-made suits, although effective in certain contexts, often suffer from issues like drift and the requirement for numerous sensors, without directly aligning reconstructions with video data. Conversely, this work introduces a less sensor-intensive method that employs a blend of 5-6 IMUs and video from a moving camera to align and optimize the 3D pose estimation process, mitigating drift and enabling more dynamic capture scenarios beyond fixed recording volumes.

This methodology leverages the SMPL[Lop15] body model and incorporates a unique video-inertial fusion technique to accurately map IMU data to 3D poses while correcting for heading drift, a common challenge in IMU-based systems. This architecture is depicted in Fig 2.11, which illustrates the comprehensive process beginning with the fitting of the SMPL body model to IMU data to derive initial 3D poses ($\hat{\Theta}$). This initial step is crucial for establishing a baseline from which more intricate pose estimations can be built. Subsequently, the method processes 2D poses (V) extracted from video footage, aiming to establish a universally consistent correlation between

2D and 3D poses. A pivotal aspect of this methodology is the simultaneous optimization of camera poses (Ψ), heading angles (Γ), and 3D poses (Θ) against the backdrop of IMU and video data integration. This multifaceted optimization not only ensures the precision of pose estimation but also compensates for potential drifts and inaccuracies typically associated with IMU data. In an iterative refinement phase, the model incorporates feedback on camera poses and heading angles, thereby enhancing the accuracy and reliability of pose assignment and tracking outcomes.

The approach not only demonstrates the feasibility of capturing realistic human motion with fewer sensors but also showcases the potential for application in diverse and challenging real-world settings. Through a detailed comparison with existing datasets and methodologies, the paper underscores the limitations of current datasets primarily confined to indoor or controlled settings and highlights the complementary nature of their proposed dataset, 3DPW, which introduces more varied and challenging sequences. This work stands out for its innovative use of technology to bridge the gap between the fidelity of MoCap in studio settings and the versatility required for in-the-wild applications, thus offering a valuable perspective for researchers and practitioners in the field of pose estimation.

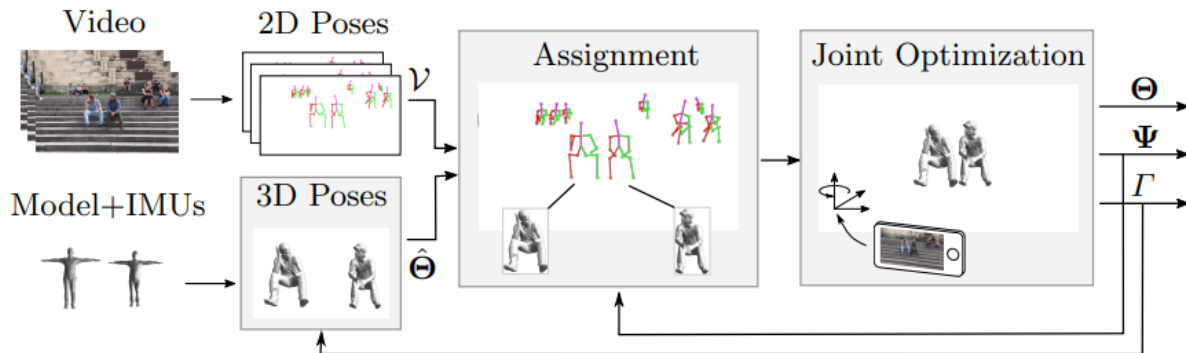


Figure 2.11: Overview of the Pose Estimation Pipeline. This diagram illustrates the initial acquisition of 3D poses through the application of the SMPL body model to orientations measured by IMUs. Following this, the system seeks to identify and assign a consistent match between these initial 3D poses and the 2D poses detected in video footage. The pipeline emphasizes the joint optimization of the 3D poses, camera orientations, and heading adjustments, utilizing feedback from the initial data to refine and enhance the accuracy of pose estimation and tracking. This iterative process allows for precise alignment and synchronization of IMU and video data, showcasing the model’s innovative approach to improving real-world motion capture.

Following the comprehensive literature review of 3D pose estimation models, Table 2.1 presents a detailed comparison of various models, highlighting their Average MPJPE (mm), usage of 2D

ground-truth joints, perspective (multi-view or monocular), and the year of publication. This comparison underscores the evolving accuracy and methodology trends within the field.

Dataset	Model	Average MPJPE (mm)	Multi-View or Monocular	Year
Human3.6M	[Chu23a]	15.6	Multi-View	2023
Human3.6M	[Zhu23]	16.9	Monocular	2022
Human3.6M	[Chu23b]	17.6	Multi-View	2022
Human3.6M	[Isk19]	17.7	Multi-View	2019
Human3.6M	[Red21]	18.7	Multi-View	2021
Human3.6M	[He20]	19.0	Multi-View	2020
Human3.6M	[Zha20]	19.5	Multi-View	2020
Human3.6M	[Sha23]	19.6	Monocular	2023
Human3.6M	[Meh23]	38.4	Monocular	2023
TotalCapture	[Von18]	26.0	Monocular + IMUs	2018

Table 2.1: Comparison of 3D Pose Estimation Models

2.3 Evaluation Metrics

2.3.1 2D Human Pose Estimation

Evaluating the accuracy and effectiveness of 2D HPE models necessitates a variety of metrics due to the complexity of the task, which can involve varying human body sizes, poses, and the number of subjects in an image. This section delineates the predominant metrics that have been adopted in the literature for assessing 2D HPE models.

Percentage of Correct Parts (PCP)

initially served as a principal metric for evaluating the precision of limb predictions in 2D HPE. It assesses the accuracy with which models could localize limbs compared to ground truth annotations. The criterion for a limb’s correct localization is expressed by the following equation, where the distance between the model’s predicted joint positions and the actual joint positions must not exceed a predefined fraction of the limb’s length, typically set between 0.1 and 0.5 of the limb length [Eic12].

$$\text{PCP} = \frac{\sum_{i=1}^N \mathbb{I}\left(\frac{\|p_i - gt_i\|}{l_i} < \theta\right)}{N}, \quad (2.1)$$

where θ is the threshold for considering a prediction correct. The PCP metric, especially $PCP@0.5$, indicates that a model’s performance is superior as the PCP value increases.

Percentage of Detected Joints (PDJ)

was developed to address some of the limitations associated with PCP, focusing on the proportion of accurately predicted joints. A joint is considered accurately predicted if its distance from the true joint position is within a specific fraction of the torso’s diameter, as delineated below [Tos14].

$$PDJ = \frac{\sum_{i=1}^N \mathbb{I}(\|p_i - gt_i\| < \alpha \cdot d_{torso})}{N}, \quad (2.2)$$

with α representing the proportional threshold related to the torso diameter.

Percentage of Correct Keypoints (PCK)

assesses the accuracy of keypoint localization within a specified threshold. This threshold is commonly set to a proportion of the head segment length or the torso diameter, making PCK a versatile metric for evaluating model performance [Yan12].

$$PCK = \frac{\sum_{i=1}^N \mathbb{I}(\|p_i - gt_i\| < \beta \cdot l_{head})}{N}, \quad (2.3)$$

where β defines the threshold as a proportion of the head segment length, *e.g.*, $PCKh@0.5$.

Object Keypoint Similarity (OKS)

is adapted for keypoint evaluation, similar to the Intersection over Union (IoU) metric used in object detection. OKS accounts for the scale of the subject and the accuracy of the predicted keypoints in relation to the ground truth, making it an integral metric for keypoint challenge evaluations, such as those in the COCO dataset [Lin14].

$$OKS = \sum_i \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right) \delta(v_i > 0), \quad (2.4)$$

Here, d_i is the Euclidean distance between the predicted and actual keypoint positions, s represents the scale of the person, and k_i is a per-keypoint constant modulating the fall-off.

Through these metrics—PCP, PDJ, PCK, and OKS—a comprehensive framework is provided for evaluating the performance of 2D HPE models, each metric offering insights into different aspects of model accuracy in localizing human body parts.

2.3.2 3D Human Pose Estimation

Evaluating the performance of 3D HPE models introduces additional complexity over its 2D counterpart, necessitating metrics that accurately reflect the depth and spatial accuracy of estimated poses in three-dimensional space. Unlike 2D HPE, where ground truth is often obtained from manual annotations in images, the accurate positions required for evaluating 3D HPE models typically come from MoCap systems. These systems use multiple cameras or sensors to track reflective markers placed on a subject's body, providing precise 3D coordinates for each marker. This MoCap data serves as the ground truth against which the 3D poses estimated by computational models are compared. The primary metrics used in the literature for assessing 3D HPE models include:

MPJPE

stands as the cornerstone metric for evaluating 3D HPE models, offering a direct measure of accuracy by computing the Euclidean distance between the estimated and actual positions of joints in three-dimensional space. The formula for calculating MPJPE is expressed as:

$$MPJPE = \frac{1}{N} \sum_{i=1}^N \|\mathbf{J}_i - \mathbf{J}_i^*\|_2, \quad (2.5)$$

where N represents the total number of joints, \mathbf{J}_i denotes the ground truth position of the i^{th} joint, and \mathbf{J}_i^* signifies the estimated position of the i^{th} joint. A lower MPJPE value indicates a higher precision in pose estimation, making it a fundamental metric for model performance evaluation.

PA-MPJPE

or *Reconstruction Error*, further refines the MPJPE by applying a Procrustes analysis to align the estimated pose with the ground truth before error computation. This adjustment allows for the exclusion of global position and orientation discrepancies, focusing solely on the pose accuracy:

$$PA - MPJPE = \frac{1}{N} \sum_{i=1}^N \|\mathbf{J}_{i,aligned} - \mathbf{J}_i^*\|_2, \quad (2.6)$$

where $\mathbf{J}_{i,aligned}$ represents the estimated joint positions after alignment. This metric is particularly useful for applications where the relative positioning of joints is more critical than their absolute locations in space.

Normalized Mean Per Joint Position Error (NMPJPE)

introduces a normalization step to the MPJPE calculation, scaling the predicted pose to match the size of the ground truth pose. This normalization helps mitigate the impact of size variance among subjects, ensuring the metric focuses on pose accuracy irrespective of the individual’s dimensions [Rho18].

Mean Per Vertex Error (MPVE)

extends the evaluation from joints to the entire mesh by measuring the Euclidean distance between predicted and actual vertices of the 3D model [Pav18]:

$$MPVE = \frac{1}{N} \sum_{i=1}^N \|\mathbf{V}_i - \mathbf{V}_i^*\|_2, \quad (2.7)$$

where N denotes the total number of vertices, \mathbf{V}_i is the ground truth position of the i^{th} vertex, and \mathbf{V}_i^* represents the estimated position. MPVE provides a comprehensive assessment of the model’s accuracy in reconstructing the human form, making it invaluable for applications requiring detailed body shapes.

3D Percentage of Correct Keypoints (3DPCK)

adapts the PCK metric for 3D space, evaluating the proportion of estimated joints falling within a predefined distance threshold from their true positions. This metric emphasizes the accuracy of joint localization in a three-dimensional context, accommodating the intricacies of depth perception and spatial positioning:

$$3DPCK = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\|\mathbf{J}_i - \mathbf{J}_i^*\|_2 < \tau), \quad (2.8)$$

where τ is the chosen distance threshold, and \mathbb{I} denotes the indicator function, returning 1 when the condition is met and 0 otherwise. This metric is especially pertinent for applications where precise joint localization in 3D is paramount.

The comprehensive review of current methodologies in human pose estimation, particularly within the context of sports analytics, underscores significant advancements enabled by deep learning. Techniques such as RTMPose, DWPose, and innovations in multi-view and IMU-assisted 3D pose estimation have markedly improved the precision and applicability of pose estimation technologies. However, despite these advancements, several critical areas remain underexplored,

particularly in adapting these technologies to dynamic and complex environments like water sports, including surfing.

Firstly, existing studies predominantly focus on controlled environments or scenarios with minimal background variability, limiting their applicability to outdoor sports settings, where environmental factors introduce significant challenges. Moreover, while multi-view and IMU-assisted methodologies offer promising directions, their integration and optimization for sports analytics, especially in uncontrolled environments, are not extensively explored. The nuances of water sports, characterized by rapid, unpredictable movements and unique interaction dynamics with the environment, pose additional challenges that are yet to be fully addressed. Specifically, the accuracy and robustness of pose estimation in the presence of occlusions, water splashes, and variable lighting conditions remain areas for enhancement.

Furthermore, the literature reveals a gap in the exploration of tailored methodologies that leverage both video and IMU data for enhanced pose estimation in sports. While some studies have begun to integrate these data sources, there's a lack of comprehensive frameworks that efficiently fuse this information to improve estimation accuracy and reliability, particularly for the complex dynamics of surfing. Given the identified gaps, this study aims to develop and validate an advanced pose estimation framework tailored to the dynamic and challenging environment of surfing.

Chapter 3

Methods

Building on the identified gaps in the literature, particularly the need for enhanced pose estimation techniques in dynamic sports environments, this study focuses on the unique challenges presented by surfing on artificial river waves. Surfing, especially in controlled environments like artificial waves, provides a novel context for applying and testing advanced pose estimation methodologies. Unlike ocean waves, artificial river waves offer consistent, repeatable conditions, making them ideal for systematic data collection and analysis.

The "Fuchslochwelle" in Nuremberg represents a prime example of such an environment. This artificial wave, generated in the Eibach district on the Pegnitz river, utilizes a submerged construction to create a standing wave. This setup not only attracts surfing enthusiasts but also presents a unique opportunity for sports technology research [24b]. The consistent wave conditions provided by the Fuchslochwelle contrast with the variable and unpredictable nature of ocean waves. This consistency enables a focused investigation into the nuances of human movement and pose estimation in surfing, free from the confounding variables typical of natural environments.

This study aims to leverage the unique environment provided by the Fuchslochwelle to address the research gap identified in the literature: the need for robust pose estimation methodologies capable of handling the complexities of water sports. By focusing on surfing in an artificial river wave setting, we intend to develop a pose estimation framework that not only improves accuracy and reliability in dynamic sports environments but also contributes to the broader field of human motion analysis.

To achieve this, our methodological approach encompasses several key steps:

1. **Data Collection:** Synchronized multi-view video and IMU data will be collected from surfers navigating the Fuchslochwelle. This setup is designed to capture a comprehensive dataset of surfing movements under controlled wave conditions.

2. **Data Processing:** Advanced processing techniques will be applied to ensure effective synchronization and integration of video and IMU data, addressing a critical need identified in the literature for high-fidelity data fusion.
3. **Model Development and Validation:** Based on the processed dataset, we will adapt and refine existing deep learning models for pose estimation, tailoring them to the specific challenges of surfing on artificial river waves. The performance of these models will be rigorously evaluated against established benchmarks and the unique requirements of the surfing context.

By systematically addressing these methodological aspects, the study will fill a significant gap in the existing research landscape, offering new insights into pose estimation technologies and their application in dynamic and complex sports environments.

3.1 Data Collection

The data collection for this study was conducted using a comprehensive protocol designed to ensure the consistency and reliability of the data gathered. This protocol consisted of several stages, each with a dedicated set of tasks and checks. Furthermore, it is noteworthy to mention that the protocol received full approval from the Ethics Committee of the FAU, under proposal number 22-437-B.

The initial preparation involved setting up the necessary equipment, ensuring that all devices were fully charged and functioning correctly. This included two cameras, two iPhones, two iPads, a GoPro, additional batteries, two camera tripods, two phone tripods, a GoPro suction cup mount, and a GoPro safety string. Ten Vicon (New Zealand) IMUs were also prepared by charging and labeling them, and calibration checkerboards were set up.

On the day of the measurement, the video cameras were turned on, aligned with the tripods, and set to the same frame rate and resolution. The Vicon Blue Trident sensors were connected and configured to High G (up to 200g accelerometer range), $\pm 2000^\circ/\text{sec}$, with video recording and logging mode on. All regions of interest were defined, and calibration signs were checked and adjusted as necessary.

Before the measurement, the participants were asked to sign a consent form and provide personal information such as age, height, shoe size, front leg, and wetsuit size. To facilitate the attachment of IMUs to the participants' bodies, pockets were specifically sewed onto the surfers' wetsuits and neoprene shoes. These modifications ensured a secure and unobtrusive placement of the IMUs at various locations on the participant's body, as detailed in Table 3.1.

Segment	Origin	Sensor	Position in cm		
			x	y	z
Right Femur	Mid of upper thigh, lateral	-	-	-	-
Right Tibia	Mid of lower thigh, lateral	-	-	-	-
Right Foot	Right Shoe Pocket	-	-	-	-
Left Femur	Mid of upper thigh, lateral	-	-	-	-
Left Tibia	Mid of lower thigh, lateral	-	-	-	-
Left Foot	Left Shoe Pocket	-	-	-	-
Pelvis	Directly on L1	-	-	-	-
Right Humerus	Mid of upper arm, lateral	-	-	-	-
Right Ulna	Mid of lower arm, lateral in NNPosition	-	-	-	-
Left Humerus	Mid of upper arm, lateral	-	-	-	-
Left Ulna	Mid of lower arm, lateral in NNPosition	-	-	-	-
Torso	Upper back	-	-	-	-

Table 3.1: IMUs attachment position table of the protocol

Surfboard Measures

The surfboard's vertical and horizontal axis lengths, approximate rear and front foot positions from the center, and height were measured and recorded, as illustrated in Fig.3.1.

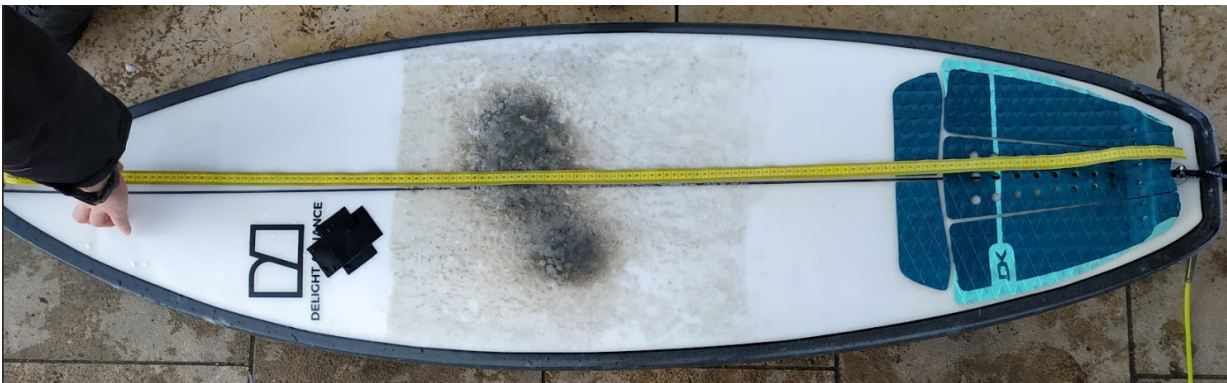


Figure 3.1: Surfboard picture with a measurement tape: This image showcases the methodical measurement of the surfboard's dimensions, illustrating the precise length of the board as captured by the measurement tape.

Calibration Movements for IMU

The participants were then asked to perform a series of calibration movements, each performed five times. These movements included N-Pose, bowing, leg up and down from the hip, foot up and down from the ankle, and arm forward and backwards, as detailed in Table 3.2. The calibration of the IMUs was meticulously conducted using the Ferraris procedure, a method that does not require angular velocity standards for calibration [Fer94]. This procedure is instrumental in ensuring the accurate alignment of the sensor orientation with the corresponding body segment orientation, thereby enhancing the precision of the IMU data collected.

Movement	File (TrialXXX)
N-Pose (Shoulder width, one foot per FP, let arms hang)	-
Bowing (straight back)	-
R: Leg up and down from hip (do not bend knee, upper body stable)	-
L: Leg up and down from hip (do not bend knee, upper body stable)	-
R: Foot up and down from ankle	-
L: Foot up and down from ankle	-
R: Arm for- and backwards (Palms to the body, not backwards)	-
L: Arm for- and backwards (Palms to the body, not backwards)	-

Table 3.2: Protocol calibration Movements for IMUs

Camera Setup

The camera setup was meticulously arranged to capture the participant from multiple perspectives during the surfing activity. This setup utilized both cameras and smartphones, strategically positioned to ensure comprehensive coverage from four distinct angles relative to the wave's direction, with specific angles formed with the river's borders to optimize the field of view. The equipment configuration and their angular relationships with the river's borders were as follows:

- **Front-Left:** An iPhone positioned closest to the surfer on the left side of the river, forming approximately a 60-degree angle with the left river border. This device was aimed towards the center of the surfing zone.
- **Back-Left:** A camera placed further back on the left side of the river, making around a 30-degree angle with the left river border, also focused towards the center of the surfing zone.

- Front-Right and Back-Right: The devices on the right side of the river were mirrored in placement and angles relative to those on the left, ensuring symmetrical coverage and angles of view.

This angular setup of the devices was deliberately chosen to optimize the coverage of the surfing activity, ensuring that key moments and movements could be captured from every critical angle. The positioning of each device, as well as the specific angles they formed with the river borders, were crucial in providing a holistic view of the participant's performance. Fig. 3.2 and 3.3 illustrate the positioning and field of view of each device.



Figure 3.2: Detailed View of the Camera and Smartphone Setup for Surfing Documentation

Once the setup was completed, two male participants, both aged XY and members of the local surf club, engaged in the surfing activity. Throughout the session, their performance was captured on video from multiple strategically positioned angles, enhancing the thoroughness of the analysis.

This elaborate setup involved devices placed at front-left, back-left, front-right, and back-right positions relative to the surfer's path. These placements were carefully chosen to capture a variety of surfing maneuvers, offering a rich dataset for evaluating the performance of the pose estimation models in dynamic, yet carefully monitored environments.

As the participants executed a series of surfing maneuvers, each movement was recorded from these different perspectives, ensuring a comprehensive view of their actions and interactions with the waves. Following the completion of the surfing sessions, all recording devices were promptly stopped, and the participants were thanked for their invaluable contribution to the study. The collected data, which included video recordings and sensor readings from the IMUs, were meticulously organized and archived. This systematic approach to data collection and management sets a solid foundation for the subsequent phase of detailed analysis, ensuring that the information is well-prepared for processing and evaluation.

However, this phase of the project was not without its challenges. During the data collection process, several challenges were encountered. One of the primary issues was the failure of certain camera batteries due to low temperatures. This necessitated the replacement of the back left camera with a smartphone. However, the available mount was incompatible with the smartphone in landscape mode, forcing the use of portrait mode for recording. This resulted in a lower resolution for that particular view. Additionally, issues were faced with the GoPro mount, which did not adhere securely to the board. This limited the ability to investigate pose estimation using an onboard camera.

3.2 Data Processing

Transitioning to the data processing stage is critical in our study, serving as the bridge between collecting raw data and conducting insightful analysis. This stage is tasked with transforming the collected data into a coherent and synchronized dataset, ready for detailed examination. Despite our thorough planning and execution during data collection, we faced several challenges that required adaptive and creative processing strategies. Our main goals during this phase included multi-view camera synchronization, multi-view camera calibration, and IMU-video synchronization. Each goal focused on refining specific aspects of our dataset, ensuring its integrity and reliability for the analysis that lies ahead.

3.2.1 Multi-View Camera Synchronization

The process of multi-view synchronization posed important challenges, mainly due to the environmental factors that were present during the data collection phase. The initial plan was to leverage Adobe Premiere software, a widely-used platform for video editing and post-production, to automatically synchronize the footage obtained from the four different camera views. This technique was intended to be carried out using audio cues, an efficient method of aligning video footage from multiple sources, provided the environment is relatively quiet.

Unfortunately, the environmental conditions present during our data collection were not conducive to this approach. The location of our experiment, being in close proximity to the river waves, resulted in a substantial amount of background noise. This made it difficult for the software to discern the audio cues necessary for auto-synchronization. This unexpected obstacle necessitated a reevaluation of our synchronization strategy.

To address this challenge, a more manual approach was adopted to achieve synchronization. This necessitated a meticulous examination of the footage from all four camera views to identify a specific reference frame that was visible across all videos. The selected frame was a moment when the participant was in mid-jump during the IMU calibration process. This action was chosen due to its distinctive visual appearance, making it an ideal candidate for a synchronization marker.

Utilizing Adobe Premiere, all four views were aligned with the chosen reference frame. This process, while more time-consuming and labor-intensive than the originally planned automatic synchronization, was successful in achieving accurate multi-view synchronization. This facilitated the progression to the subsequent steps of the research methodology, namely multi-view calibration and IMU-video synchronization.

3.2.2 Multi-View Camera Calibration

Multi-view calibration is a critical process in multi-view camera setups, enabling the transformation of two-dimensional image points to three-dimensional real-world points. The objective of this process is to estimate the intrinsic and extrinsic parameters of each camera in the system. The intrinsic parameters are unique to each camera and include attributes such as focal length and optical center, while the extrinsic parameters describe the position and orientation of each camera in the space [Kau24].

In the context of this study, a calibration checkerboard was utilized, a tool commonly used for camera calibration due to its geometrically regular pattern that simplifies the identification of matching points across different views. The participant held this checkerboard while standing

on the surfboard, thus providing a reference that could be seen from all camera angles [Kau24]. Alongside the handheld checkerboard, one static checkerboard was positioned on each side of the river.

The process commenced with the identification of the checkerboard pattern in the images captured by the cameras as illustrated in Fig 3.5. The corners of the squares on the checkerboard served as reference points, and their positions in the image plane were detected by the calibration algorithm [Kau24].

Following the identification of these points, the next step was to estimate the aforementioned intrinsic and extrinsic parameters. The intrinsic parameters, which account for lens distortion and other camera-specific characteristics, were first estimated using a pinhole camera model. The mathematical representation of these parameters can be represented as:

$$A = \begin{bmatrix} f_x & s & o_x & 0 & f_y & o_y & 0 & 0 & 1 \end{bmatrix} \quad (3.1)$$

where f_x and f_y are the focal lengths along the x and y axes, s is the skew coefficient, and o_x and o_y are the coordinates of the optical center.

Subsequently, the extrinsic parameters were calculated. These parameters, which include the rotation and translation vectors for each camera, describe the position and orientation of the cameras in 3D space. By knowing the real-world distances between the corners on the calibration checkerboard, a perspective transformation could be performed to deduce the spatial relationship between the cameras. The extrinsic parameters can be represented as:

$$[R|T] = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \end{bmatrix} \quad (3.2)$$

where R is the rotation matrix and T is the translation vector.

3.2.3 IMU-Video Synchronization

The synchronization of IMU data with video footage is a critical process in our methodology. The main objective of this synchronization is to see the potential of IMUs in enhancing the accuracy of 3D pose estimation, given that the IMU data offers a rich source of detailed information about the participant's motion, including aspects such as acceleration and orientation.

To facilitate this synchronization, an animated plot was generated from the IMU data. This plot provides a visual representation of the IMU readings over time, and it is particularly useful

in identifying distinct movements or actions performed by the participant.

Following the creation of the animated plot, it was overlaid onto the video footage. This superimposition resulted in a composite resource that combines both visual and motion data, thereby offering a comprehensive understanding of the participant's movements during the surfing activity. Subsequently, reference frames were manually identified within the composite resource. These frames correspond to distinctive instances during the participant's performance, such as jumps, landings, or claps. These actions are easily recognizable both visually and in the IMU data, making them ideal synchronization points.

The video footage was then aligned with the corresponding points in the animated IMU plot using these reference frames. This alignment process, while requiring a high level of detail and precision, ensured that the video and IMU data were accurately synchronized. The following step in this synchronization effort was the adjustment of the IMUs' sampling frequency. Originally recorded at 1600Hz, the IMU data was downsampled to 59.94Hz to align with the video recordings' frame rate. This specific frame rate was chosen instead of the more rounded figure of 60Hz to match the lowest refresh rate found among the four camera views, which was 59.94 frames per second (fps). This adjustment was crucial for maintaining temporal consistency across the datasets, enabling a seamless integration of the high-resolution motion data with the visual footage.

Through this process of IMU-video synchronization, the integration of the detailed motion data from the IMUs with the visual data from the video footage was achieved. This integrated dataset offers a more comprehensive basis for the subsequent 3D pose estimation, potentially improving its accuracy and robustness.

3.3 2D Pose Estimation

The process of 2D pose estimation is crucial in our research methodology, setting the stage for the subsequent phase of 3D pose estimation. This step involves identifying key body joint positions within the video footage's two-dimensional frame. Reflecting on our literature review, we have already discussed the capabilities of RTMPose and DWPose. To build on this foundation, our study also explored additional models including AlphaPose [Fan22], recommended by the 3D pose estimation model MotionBert.

The initial exploration of 2D pose estimation was conducted using AlphaPose, a widely used tool for human pose estimation, on the surfing videos in the dataset. The performance of AlphaPose was qualitatively evaluated based on its predictions on the collected dataset. It was noted

that AlphaPose encountered certain difficulties in accurately identifying the positions of the joints. There were instances where the model misinterpreted the right and left sides of the body. Moreover, it struggled to locate joint positions when faced with obstacles such as wave splashes or the surfer’s wetsuit. A specific example of AlphaPose’s inability to accurately predict the joints under these conditions is depicted in Fig. 3.7.

Given the observed limitations of AlphaPose, and to expand upon our foundation, we also explored ViTPose and RTMO [Xu22; Lu23], two models that, while not covered in our initial literature review, present cutting-edge advancements in pose estimation technologies. RTMO introduces a one-stage pose estimation framework that effectively balances speed and accuracy by integrating coordinate classification within the YOLO architecture. It employs dual 1-D heatmaps for keypoint representation, a dynamic coordinate classifier, and a tailored loss function, outperforming other one-stage models with a considerable increase in AP and speed on the COCO dataset.

ViTPose leverages the scalability and parallelism of plain vision transformers, demonstrating their unexpectedly strong capabilities in pose estimation. It uses a straightforward model structure with a scalable capacity from 100M to 1B parameters and a lightweight decoder. This flexibility allows ViTPose to set a new benchmark in the balance between throughput and performance, with its largest model establishing a new state-of-the-art in the MS COCO Keypoint Detection benchmark. The inclusion of ViTPose and RTMO in our analysis ensures a comprehensive evaluation of current technologies in 2D pose estimation, strengthening our approach towards developing robust 3D HPE. The evaluation process encompassed both quantitative and qualitative assessments. The design of the evaluation took into consideration the unique challenges presented by the dataset. For instance, scenarios involving occlusion, such as a surfer’s foot partially obscured by a wave, and low-contrast situations like a black wetsuit against the water, were contemplated.

The models were further assessed on their ability to tackle these challenges and deliver accurate predictions. This evaluation entailed a comparison of the models’ predictions against ground truths, with a particular emphasis on how well each model coped with occlusion and low-contrast scenarios. In addition to these specific assessments, a more encompassing analysis of the models’ performance was conducted by calculating the average confidence of predictions for each joint across views from different cameras. The aim of this analysis was to provide a comprehensive overview of the performance of the different models and offer insights into their overall reliability and robustness.

Through this exhaustive evaluation strategy, the intent is to pinpoint the models that best cater to the specific requirements and challenges of this study. The insights derived from these findings

will lay the foundation for the subsequent work on 3D pose estimation. The detailed results of this analysis, including the comparative performance of the selected models and their applicability to 3D pose estimation in the context of surfing, will be thoroughly discussed in the Results section.

3.4 3D Pose Estimation

3.4.1 Single-View 3D Pose Estimation

The transition from 2D pose estimation to 3D pose estimation is a significant step in this research study. It involves converting the two-dimensional positions of body joints into a three-dimensional space, thereby providing a more extensive understanding of the body's movements and positions. In the realm of 3D pose estimation, several models such as BCP [Chu23a], MotionBERT [Zhu23], LMT [Chu23b], and Learnable Triangulation of Human Pose have demonstrated high performance. However, since BCP and LMT are not open source and thus not accessible for this study, the focus has been placed on the MotionBERT and Learnable Triangulation models.

The decision to explore the MotionBERT and Learnable Triangulation models was based on a rigorous review of the literature. As depicted in the ranking plot in Fig. 3.8, these models stand as the next best alternatives following BCP and LMT. Their reported high performance in 3D pose estimation tasks on the Human3.6M dataset underscores their potential suitability for this study. In the context of estimating 3D joint positions, the reprojection error serves as an efficient evaluation metric, particularly due to the absence of ground truth 3D positions. This metric, discussed in the Literature Review, consequently fills the gap by providing a measure of alignment between the original 2D poses and their corresponding 3D estimations, once these are reprojected back onto the 2D plane.

This error, essentially, quantifies the discrepancy between the input 2D pose, denoted by \mathbf{T} , and the predicted 3D positions of the joints, symbolised by \mathbf{P} . The predicted 2D positions are represented by \mathbf{P}' . The confidence in the predictions is captured by \mathbf{C} . The difference, scaled by the confidence and denoted by \mathbf{D} , is calculated as $\mathbf{D} = (\mathbf{P}' - \mathbf{T}) \times \mathbf{C}$.

The reprojection error, a key factor in determining the accuracy of the 3D pose estimations, is then derived as the mean of the L2 norm (Euclidean norm) of this difference. This can be mathematically expressed as:

$$\text{Reprojection Error} = \text{mean}(\|\mathbf{D}\|_2) \quad (3.3)$$

A smaller reprojection error indicates a strong alignment between the predicted 3D positions

and the actual 2D poses, thus implying a higher accuracy of the model. This metric therefore plays a fundamental role in assessing the performance and validity of the 3D pose estimation model.

The findings from the 3D pose estimation phase will complement the insights derived from the 2D pose estimation phase. With a comprehensive understanding of the body's movements and poses in both two and three dimensions, a more holistic approach to pose estimation can be achieved. The knowledge gained from these evaluations will serve as the foundation for further research and applications in the field of human pose estimation.

3.4.2 Exploration of Single-View 3D Pose Prediction Enhancement Using The Surf Pose Data

The objective of this research phase is to enhance the 3D pose predictions on our acquired surf dataset by fine-tuning the model to reduce the reprojection error. The proposed model will undergo joint training, utilizing both the Human3.6M [Ion14] dataset and the surf dataset we have collected. This dual-dataset approach aims to minimize the MPJPE on the Human3.6M dataset, while also focusing on reducing the reprojection error within our surf dataset. The strategy for this tailored training regimen is illustrated in Fig. 3.9.

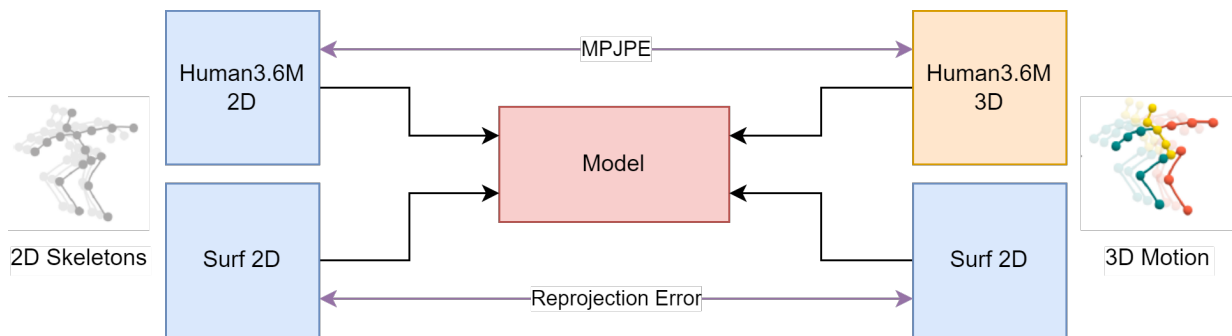


Figure 3.9: The proposed strategy for joint training on the Human3.6M and surf datasets, focusing on fine-tuning the MotionBERT model (represented by the red block). This process is aimed at improving the accuracy and reducing reprojection errors in 3D pose estimations, specifically tailored for surfing activities. The blue blocks represent 2D data inputs, while the orange block signifies 3D data inputs. The purple lines depict the loss pathways, illustrating the feedback mechanism used to adjust MotionBERT's parameters during the fine-tuning process.

The reprojection error, as previously outlined, evaluates the degree of alignment between the estimated 3D poses and the actual 2D poses once reprojected onto the 2D plane. In this context,

a reduction in the reprojection error signifies an improvement in the precision of the model's 3D pose estimations. Therefore, the reprojection error will serve as a primary performance metric in the evaluation of the model's efficacy.

In parallel with efforts to minimize reprojection error, the model also aims to reduce the MPJPE, a metric already introduced as crucial for evaluating 3D pose estimation accuracy. By measuring the average Euclidean distance between the actual and predicted joint positions in 3D space, a lower MPJPE value reflects improved accuracy and performance of the model. The Human3.6M dataset, chosen for its status as the largest and most widely accepted benchmark in human pose estimation, provides a comprehensive basis for these evaluations.

The attempt to minimize these errors on both datasets concurrently represents a novel approach to the training of 3D pose estimation models. The insights gained from this two-pronged approach will provide valuable contributions to the field of human pose estimation, particularly in scenarios where ground truth 3D poses are unavailable. This approach will also enable the assessment of the model's performance under a more diverse set of conditions, thereby providing a more robust evaluation of its capabilities.

The results from this phase of the research will not only supplement the findings from the 2D pose estimation phase, but will also contribute to a more comprehensive understanding of the body's movements and poses in three dimensions. This understanding will be instrumental in advancing subsequent research and applications in human pose estimation.

3.4.3 Exploration of Single-View 3D Pose Prediction Enhancement Through IMU Integration

This phase of the research aims to explore the potential benefits of incorporating IMU data into the 3D pose prediction model. By combining the surf dataset with synchronized IMU data, the study seeks to understand if the additional motion context provided by IMUs can lead to an enhanced estimation of 3D poses. Unlike the solely 2D pose-based approach discussed previously, this methodology introduces a combination of 2D pose information and features extracted from IMU readings. Specifically, the model's input configuration is expanded from solely 2D pose information with a shape of (batch_size, number_of_frames, 17, 3) to include a new dimension derived from the IMU data. Here, "17" corresponds to the number of keypoints representing human joints, "3" indicates the three dimensions of each keypoint (x-axis, y-axis, and the confidence level of each keypoint's detection), and "number_of_frames" represents the temporal dimension accounting for the sequence of frames within a batch.

The integration process employs an embedding layer, essentially a single-layer neural network,

designed to process the IMU data. With raw IMU data shaped $(\text{batch_size}, \text{number_of_frames}, 60)$, the layer outputs a transformed shape that aligns with the per-frame joint keypoint representation, $(\text{batch_size}, \text{number_of_frames}, 17, 3)$. This output is then concatenated with the 2D pose data, resulting in an augmented input shape of $(\text{batch_size}, \text{number_of_frames}, 17, 6)$, where the fourth dimension represents the enriched feature set incorporating both pose and IMU data. Fig. 3.10 visually depicts this process, illustrating how the IMU data, once processed through the embedding layer, is combined with 2D pose information to serve as the enhanced input for the model.

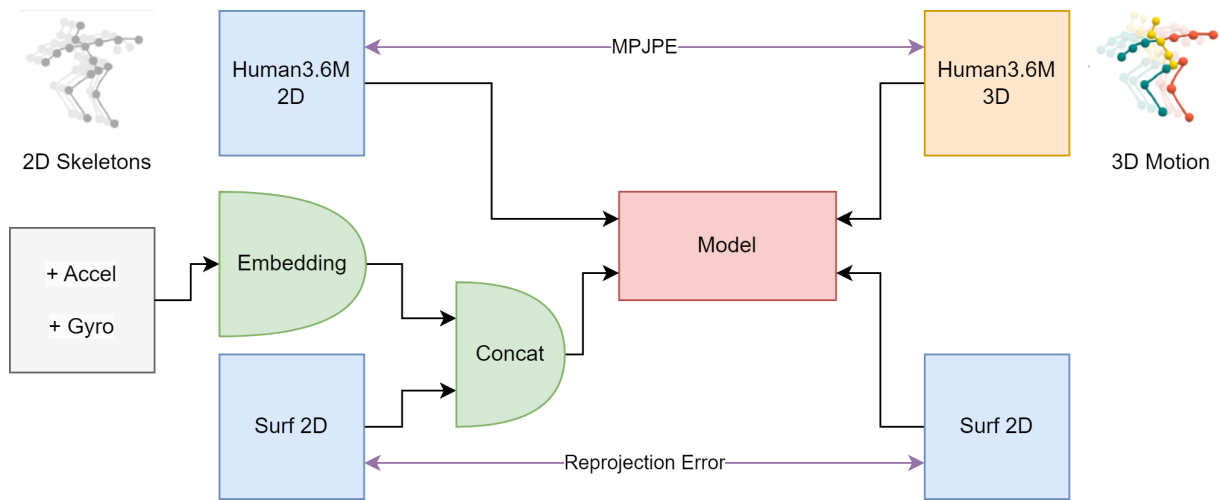


Figure 3.10: Illustration of integrating IMU data with 2D pose inputs to potentially enhance 3D pose predictions. The embedding layer processes the IMU data, which is then concatenated with 2D pose data, forming the additional model input. The embedding and concatenation operations are represented with green shapes

The exploration continues with the dual dataset training strategy, aiming to minimize the MPJPE on the Human3.6M dataset alongside the reprojection error on the surf dataset, now enriched with IMU data. This enrichment is hypothesized to offer a broader understanding of human motion dynamics, potentially improving the model's accuracy in reconstructing 3D poses.

The hypothesis is that additional contextual information about the participant's movements, provided by IMU readings, may contribute to a more accurate estimation of 3D poses. This exploration is aimed at assessing whether the augmented data set will lead to considerable improvements in the model's performance, as evidenced by reductions in both the MPJPE on the Human3.6M dataset and the reprojection error on the augmented surf dataset. The outcome of this innovative approach will offer valuable insights into the efficacy of multimodal data integra-

tion for refining 3D pose estimation models.

3.4.4 Multi-view 3D Pose Estimation

To ascertain whether monocular 3D pose estimation suffices for accurate pose predictions, this research will contrast it against a multi-view approach. By doing so, the research aims to establish a more reliable and comprehensive framework for 3D pose estimation that could overcome the constraints faced by purely monocular methods.

In the process of multi-view 3D pose estimation, the learnable triangulation model was initially applied using PoseResNet as a backbone for extracting 2D joints. This approach, however, did not yield satisfactory results, prompting a transition to the RTMO model for the 2D pose estimation phase. The RTMO model was chosen for its robust performance and was applied to two views specifically - the back-left and back-right cameras. These cameras were selected due to their high frequency of detected checkerboard patterns, a critical factor for accurate calibration.

Following calibration, the images were cropped around the surfer and then resized, an important preprocessing step that allows the model to focus on the area of interest.

The cropping and resizing of images are essential preprocessing steps in the model. However, these operations alter the spatial dimensions of the image, necessitating an update to the camera parameters to maintain their accuracy.

When an image is cropped, the principal point of the camera, represented by the coordinates (cx, cy) in the camera matrix K , changes its position. This is because the principal point is the projection center from the top left corner of the image. The new principal point coordinates (new_cx, new_cy) are calculated by subtracting the left and upper coordinates of the crop rectangle from the original cx and cy , respectively. The camera matrix K is then updated with these new values.

Resizing an image also requires an update to the camera parameters. The focal lengths (fx, fy) and the principal point coordinates (cx, cy) in the camera matrix K are scaled by the ratio of the new image size to the original image size. The new focal lengths (new_fx, new_fy) and the new principal point coordinates (new_cx, new_cy) are calculated by multiplying the original values by the ratio of the new width to the original width and the new height to the original height, respectively. The camera matrix K is then updated with these new values.

These updates to the camera parameters ensure that the spatial understanding of the image by the model remains accurate, even after the image has been cropped and resized. This is crucial for the accurate prediction of the 3D pose. The calibrated camera parameters and the cropped images were subsequently input into the learnable triangulation model, which then predicted the 3D pose.

The triangulation is done using a process that takes a sequence of projection matrices and corresponding 2D points as inputs and operates by constructing a matrix A . This matrix is formed by scaling and subtracting the projection matrices with the 2D points and their associated confidences. If no confidences are provided, all points are assumed to have a confidence of 1.0.

Using singular value decomposition (SVD) on the matrix A , the last column of the right singular vectors is selected, providing the homogeneous coordinates of the triangulated point. These homogeneous coordinates are then converted to Euclidean coordinates, yielding the final 3D coordinates of the point.

Following the prediction of the 3D pose, the reprojection error is calculated. This error is a quantitative measure of how close the reprojected points are to the original 2D points, and thus serves as an assessment of the accuracy of the 3D pose estimation using this multi-view approach.



Figure 3.3: Satellite view of the camera setup, illustrating the positioning and orientation of the four cameras used for the surfing activity data collection. The orientations of the cameras are represented by red arrows, each pointing in the direction of the camera's field of view. This setup was meticulously designed to ensure comprehensive coverage of the surfing area from multiple perspectives, thereby facilitating a robust data collection for pose estimation analysis.



Figure 3.4: Synchronized multi-view camera views depicting a single frame captured from each of the four camera perspectives of a surfer navigating a wave, illustrating the synchronization achieved across different viewpoints.



Figure 3.5: Checkerboard corners detection from the back-left and back-right camera views

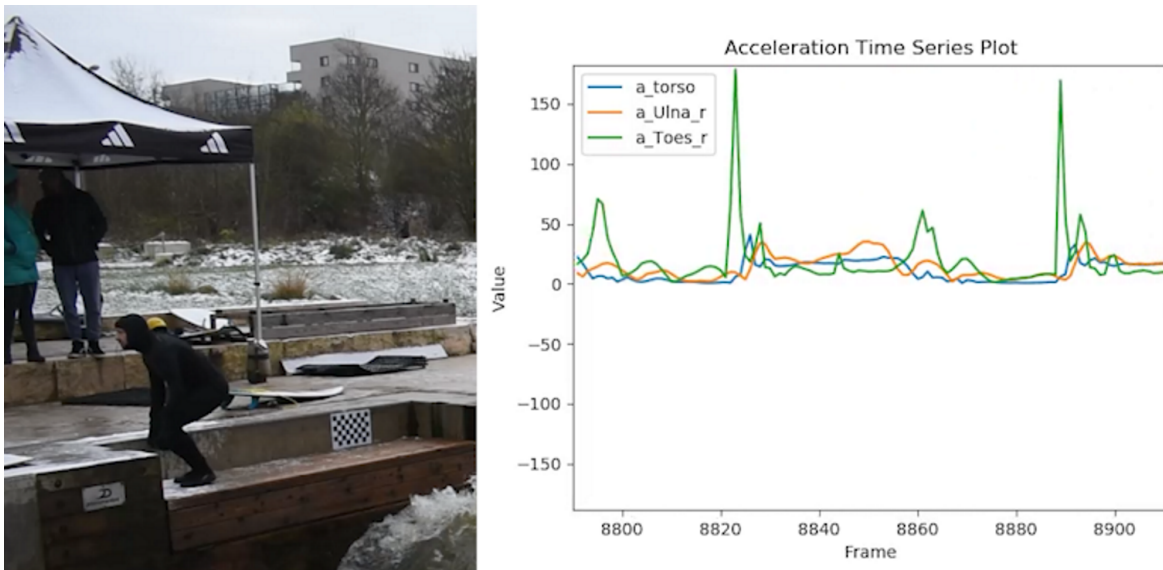


Figure 3.6: Overlay of the animated IMU data plot on the video footage, illustrating the synchronization process



Figure 3.7: Illustration of a failure case of AlphaPose in which the model was unable to correctly identify all the joints of the surfer in a specific frame. This example highlights a limitation of the AlphaPose model when faced with complex scenarios typical in dynamic sports environments, such as surfing.

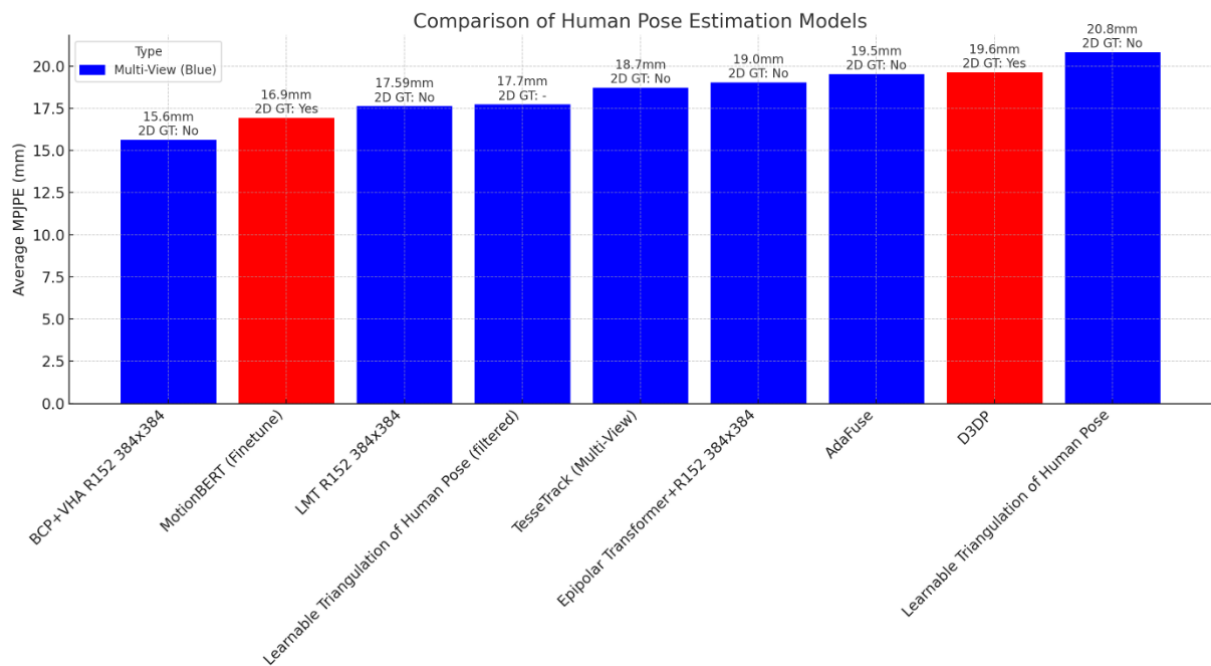


Figure 3.8: Comparative Analysis of 3D Pose Estimation Models on the Human3.6M Dataset. This ranking is based on the models’ performance, measured by the average MPJPE in millimeters (mm). Red bars represent models that utilize a monocular input approach. In contrast, blue bars denote models that employ a multi-view approach. The data used for this evaluation is sourced from <https://paperswithcode.com/sota/3d-human-pose-estimation-on-human36m>. [Chu23a; Zhu23; Chu23b; Isk19; Red21; He20; Zha20; Sha23]

Chapter 4

Results

This chapter presents the results obtained from the comprehensive analysis conducted on the surfing data collected and processed as detailed in the preceding chapters. Following the methodologies outlined in Chapter 3, both 2D and 3D pose estimations were performed, leveraging a variety of advanced computational models and techniques. Additionally, the integration of IMU data with video footage provided a rich, multimodal dataset for an enhanced understanding of dynamic surfing maneuvers.

The chapter is structured as follows. Section 4.1 showcases the findings from the 2D pose estimation, highlighting the accuracy and limitations of the models applied. Section 4.2 delves into the 3D pose estimation results, including the reprojection errors and a comparative analysis of the monocular and multi-view approaches. Subsequent sections present the results of the integration of IMU data, emphasizing the possible value added by this additional modality.

These results are not only pivotal in the advancement of technical training protocols for surfing but also contribute to the broader field of human pose estimation. The precision and reliability of the estimations are critically evaluated, demonstrating the robustness and adaptability of the proposed models to the challenging and fluid dynamics of water sports.

4.1 2D Pose Estimation Results

The quantitative analysis of 2D pose estimation models on our surf dataset is depicted through Fig. 4.1, which illustrates the average confidence scores with corresponding standard deviations as error bars. These plots quantify the models' confidence in their pose predictions, where the confidence score is derived from the output logits of the model. Mathematically, the average confidence score for a particular joint across all frames is computed by taking the mean of the

confidence levels assigned to that keypoint in each frame. Specifically, if a model outputs a confidence value c_{ij} for the j -th keypoint in the i -th frame, and there are m frames, the average confidence for the j -th keypoint is given by $\frac{1}{m} \sum_{i=1}^m c_{ij}$. This approach ensures that the metric reflects the model’s consistent performance in terms of confidence across the temporal dimension of the dataset, thereby providing insight into the model’s capability to reliably estimate poses under challenging conditions such as poor lighting and occlusions that are common in river wave surfing scenarios.

In the individual plots corresponding to each camera view, the x-axis represents the joint names, while the y-axis denotes the confidence scores. Across the four camera views, ViTPose and RTMO exhibited closely matched performance, with confidence levels generally ranging between 80-90%. Notably, RTMO displayed a decrease in confidence for the joints pertaining to the nose, eyes, and ears. This decrease was accompanied by a high variance in confidence, indicating that the confidence levels for these particular joints were exceptionally low in some cycles. Such variance suggests that RTMO may have difficulties in consistently detecting these facial features across all conditions present in the dataset. Conversely, DWPose generally outperformed RTMPose, except in the case of the back-left camera view, where their performances were quite similar.

This performance contrast is well-captured in the provided table 4.1, which consolidates the average confidences (\pm standard deviation) for all joints and non-face joints across the models. Remarkably, RTMO not only exhibited a high average confidence across non-face joints (0.954 ± 0.040) but also a notably higher variance for all joints (0.853 ± 0.051), aligning with the observed difficulty in detecting facial features. Meanwhile, ViTPose maintained high confidence levels with relatively low variance, indicating robust performance across all joints. DWPose and RTMPose offered competitive confidence scores, yet DWPose generally held an edge, especially in non-facial joints, emphasizing its effective performance except in specific camera views where it closely matched RTMPose. This data underscores the nuanced capabilities and limitations of each model within the context of joint detection accuracy and reliability.

Model	All Joints (Avg \pm Std)	Non-Face Joints (Avg \pm Std)
RTMPose	0.700 ± 0.023	0.677 ± 0.019
DWPose	0.782 ± 0.021	0.767 ± 0.021
ViTPose	0.877 ± 0.033	0.872 ± 0.028
RTMO	0.853 ± 0.051	0.954 ± 0.040

Table 4.1: Overview of average confidences (\pm standard deviation) across all joints and non-face joints

In addition to the quantitative evaluation, a qualitative comparison was conducted for two particularly challenging scenarios observed in the dataset because that gives a more comprehensive explanation of where and why errors may occur. The first scenario involves the occlusion of the surfer’s rear foot by a wave, and the second scenario features low contrast, complicating the visibility of the surfer’s front hand. For these scenarios, the 2D keypoints estimations by the different models are visually compared in 4.2.

Building on the comprehensive analysis outlined above, we further elucidate the models’ capabilities through a direct qualitative comparison, as presented in Table 4.2. This comparison specifically addresses two challenging scenarios identified within our dataset: one involving the occlusion of the surfer’s rear foot by a wave, and another characterized by low contrast affecting the visibility of the surfer’s front hand. The models evaluated, were chosen based on their pretrained versions available from MMPose [Con20], a comprehensive open-source toolbox for pose estimation. This selection allows us to leverage the advanced capabilities and pre-existing training of these models to address the unique challenges presented by our surf dataset.

By presenting the 2D keypoints estimations for each model side-by-side, the table offers an insightful examination of how each model navigates the complexities of occlusion and low contrast. It’s important to note that the keypoint sets used for each model’s evaluation may vary. This variance stems from the models’ original training configurations within MMPose, where each was trained on specific datasets with particular sets of keypoints. As such, the inference performed on our surf dataset utilized the applicable keypoint set for each pretrained model, ensuring the most accurate and relevant comparison under the conditions tested.

This qualitative comparison complements the quantitative analysis by providing a visual representation of the models’ capabilities to handle intricate situations, thereby offering a more holistic evaluation of their performance in real-world conditions.

4.2 3D Pose Estimation Results

The evaluation of the pretrained MotionBERT model’s precision and reliability was conducted through an analysis of joint angles during a walking movement, leveraging data from the Human3.6M dataset. This focus was chosen due to the critical role that elbows and knees play in representing dynamic human activities. Figure 4.2 depicts the variation in angular positions of the left and right elbows, alongside the left and right knees, over a series of frames. This visual comparison with ground truth angles offers insight into the model’s performance.

The evaluation of the MotionBERT model, pretrained as described by Zhu et al., focused on

its precision and reliability in estimating joint angles during a walking movement. This analysis utilized data from the Human3.6M dataset [Ion14], with a particular emphasis on the angles of the elbows and knees, as joints usually have the most range of motion in highly dynamic movements. Figure 4.2 illustrates the variation in angular positions of the left and right elbows, alongside the left and right knees, across a series of frames. This visual comparison is made against ground truth angles obtained from optical motion capture, offers insight into the model’s performance.

Analyzing the temporal patterns of these joint angles is vital for assessing the model’s ability to accurately capture human motion. While the congruence between the predicted and actual angles for the knees and left elbow underscores the model’s effectiveness, the slight deviation observed for the right elbow—though more pronounced—still falls within a satisfactory range, indicating a few degrees of difference. Such disparities are important to consider as they could influence the model’s applicability in detailed movement analysis, particularly in scenarios requiring precise motion tracking for athletic training or performance enhancement. The findings suggest that the MotionBERT model demonstrates substantial potential in capturing complex movements with high fidelity.

Following this initial examination, the focus shifted to the model’s 3D pose estimation capabilities. The reprojection errors for each joint across all camera views were meticulously calculated and visualized through a series of plots. These plots offer a direct comparison between the original and finetuned versions of the MotionBERT model, providing an evaluation basis for the adjustments made during the finetuning process. Figure 4.3 portrays a comparative analysis of the reprojection errors across the two model states. The x-axis indicates the joint names, while the y-axis quantifies the reprojection error. The analysis demonstrates that the finetuned model performs better overall, particularly in the front-right camera view, where it shows improvements across all joints. For other views, the finetuned model exhibits slight enhancements in most joints, with larger improvements noted in the estimation accuracy for the left ear and right shoulder joints.

The finetuned model has exhibited superior performance, particularly noted in the front-right camera view where improvements are seen across all joints. To quantitatively assess the finetuning impact, we analyzed the reprojection errors across different views on the collected surf dataset, both before and after finetuning. Notably, the reprojection error provides a measure of the discrepancy between the predicted 2D keypoints and their corresponding ground truth locations, with lower values indicating higher accuracy.

In our analysis, a pivotal aspect was the assessment of reprojection errors, specifically quantified both before and after fine-tuning the model. The fine-tuned adjustments brought about noteworthy enhancements in precision, particularly evidenced in the reprojection errors for all joints,

including and excluding facial keypoints. As detailed in the accompanying table 4.3, fine-tuning resulted in a reduction of average reprojection errors across all joints from $1.7\% \pm 0.9\%$ to $1.0\% \pm 0.2\%$, and similarly, for non-face joints from $1.5\% \pm 0.7\%$ to $1.0\% \pm 0.2\%$. These improvements are significant, highlighting the effectiveness of fine-tuning in refining the model's accuracy.

Further granularity in the results reveals that the lowest reprojection errors post-fine-tuning were observed for the "right_wrist" and "left_wrist" with values of 0.38% and 0.42% , respectively, in the Back-Right view, highlighting large improvements in model precision for limb extremities. Conversely, the highest error post-finetuning was found in the "left_knee" joint in the Front-Left view, with a value of 1.35% , suggesting areas where the model may still struggle, particularly in capturing lower limb movements accurately.

Comparing the average reprojection errors across all views, finetuning led to a reduction from 1.51% (no finetuning) to 0.98% (post-finetuning), indicating a considerable overall improvement in model performance. This comprehensive quantitative evaluation, presented alongside visual comparisons in Fig. 4.2, underscores the finetuned model's enhanced capability to accurately estimate poses under various challenging conditions, setting a new benchmark for 2D pose estimation in surf-related activities.

For context, the pretrained models from Zhu et al. were employed, leveraging their robust initial training on diverse datasets. The choice of keypoint sets for evaluation was guided by the pretrained models' configurations and their respective training regimes, which inherently influence the inference outcomes for applicable keypoint sets. This decision underscores the tailored approach to optimizing model performance for specific scenarios encountered in the surf dataset, particularly emphasizing the crucial role of lower limbs in capturing dynamic human activities as evidenced by existing literature.

This detailed quantitative analysis, alongside the qualitative insights, offers a holistic understanding of the finetuned model's efficacy, providing a solid foundation for future enhancements and applications in the domain of action sports analysis.

These results signify the impact of the joint training regimen, utilizing both the Human3.6M dataset and our surf dataset, on the model's ability to estimate 3D poses. As outlined in Section 3.4, the reduction in reprojection error indicates a more accurate alignment between the predicted 3D poses and the actual 2D poses, reaffirming the validity and precision of the fine-tuned 3D pose estimation model.

This comprehensive analysis not only demonstrates the advancements achieved through model fine-tuning but also reinforces the significance of the reprojection error as a reliable metric for assessing 3D pose estimation accuracy in scenarios devoid of ground truth 3D data. These find-

ings contribute profoundly to the overall aim of enhancing 3D pose predictions within the realm of human pose estimation in sports technology and possibly also in health applications.

4.2.1 Evaluating IMU Data Integration on Reprojection Error

In pursuit of refining 3D pose estimations, this study examined whether the inclusion of IMU data could diminish the reprojection error in single-view pose predictions. The MotionBERT model was subjected to a series of tests wherein IMU data was integrated alongside 2D pose information, as detailed in the methodological description in Section 3.4.

A plot depicting the evolution of reprojection errors over 60 epochs during the testing phase is presented in Figure 4.4. This graph illustrates the comparative analysis of reprojection errors for the MotionBERT model, both with and without the integration of IMU data, highlighting the influence of IMU data on the model's accuracy.

To ensure replicability and provide a foundation for future research, the following hyperparameters were meticulously chosen for the training process:

- **Epochs:** 60, to allow sufficient training depth without overfitting.
- **Checkpoint Frequency:** Every 30 epochs, enabling the evaluation of model performance at mid and end points of training.
- **Batch Size:** 10, balancing the computational load and gradient update frequency.
- **Dropout:** 0.0, indicating no dropout was used, to maintain all neural connections active throughout training.
- **Learning Rate:** 0.0005, chosen to optimize the convergence speed without skipping over minima.
- **Weight Decay:** 0.01, to regularize and prevent overfitting by penalizing large weights.
- **Learning Rate Decay:** 0.99, applied to decrease the learning rate gradually, helping the model to fine-tune adjustments as training progresses.

Contrary to the hypothesized benefits of IMU data, the results indicated no substantive improvement in the reprojection error. This outcome suggests that the additional motion context provided by IMUs did not translate into a more accurate estimation of 3D poses for this particular dataset and model configuration. Notably, both versions of the model, with and without IMU

data, exhibited comparable reprojection errors throughout the testing epochs, indicating that the IMU data did not detrimentally affect the model’s performance either.

The absence of improvement calls into question the direct utility of IMU data in the context of 3D pose estimation for surfing activities. It raises considerations about the complexities of integrating multimodal data and the challenges in translating such data into meaningful enhancements in model predictions. These findings prompt further investigation into the conditions and methods that might effectively leverage IMU data, guiding future research toward more sophisticated data fusion strategies or alternative pose estimation architectures that can better capitalize on the rich information offered by IMUs [Von18; Hua20a].

4.2.2 Comparison with Multi-view 3D Pose Estimation

An integral part of this research involved comparing the efficacy of monocular 3D pose estimation with a multi-view approach. The objective was to determine whether a multi-view strategy could provide a more accurate and reliable framework for 3D pose estimation, particularly in cases where monocular methods might face constraints due to occlusions or poor perspective angles.

The multi-view 3D pose estimation was conducted using the learnable triangulation model [Isk19], as outlined in Section 3.4. After processing the calibrated camera parameters and the preprocessed images, 3D poses were predicted and the reprojection errors were computed to assess the models’ performance. The results from this multi-view analysis were then compared to those obtained from both the pretrained and finetuned MotionBERT models.

The analysis demonstrates the finetuned MotionBERT model’s superior performance over the learnable triangulation model (LTM) in all joints, showcasing the significant advantages of incorporating the time dimension into pose estimation for improved reprojection accuracy. This is quantitatively supported by the data in Table 4.3, which highlights the finetuned MotionBERT model’s lower average errors and standard deviations. For instance, average errors for the finetuned model are notably low, at 0.9% for the nose and 0.6% for the left wrist, in stark contrast to LTM’s 1.56% and 3.27%, respectively. The consistency and reliability of the finetuned MotionBERT are further emphasized by its minimal standard deviation of 0.20%, indicating precise pose estimation.

Even without finetuning, the MotionBERT model outperforms LTM for most joints, excluding the left eye and right shoulder. The non-finetuned model achieves an average reprojection error of 1.5% for the nose and 0.9% for the left wrist, underscoring its efficacy. Conversely, LTM particularly struggles with joints susceptible to occlusion, like ankles, wrists, and elbows, with reprojection errors as high as 3.9% for the right ankle. These comparisons underscore the

robustness of MotionBERT models in pose estimation tasks, regardless of their finetuning status.

These findings highlight the potential of the time-aware approach that MotionBERT utilizes, emphasizing its ability to leverage temporal consistency even in single-view estimations. The consistent outperformance of the finetuned MotionBERT model underscores the value of joint training strategies that fuse datasets with different characteristics to improve the model's generalization and predictive accuracy.

The implications of these results are important for sports technology and pose estimation research. They suggest that for activities like surfing, where occlusions and rapid dynamic movements are common, models that can integrate time series information may provide a more robust framework for accurate pose estimation.

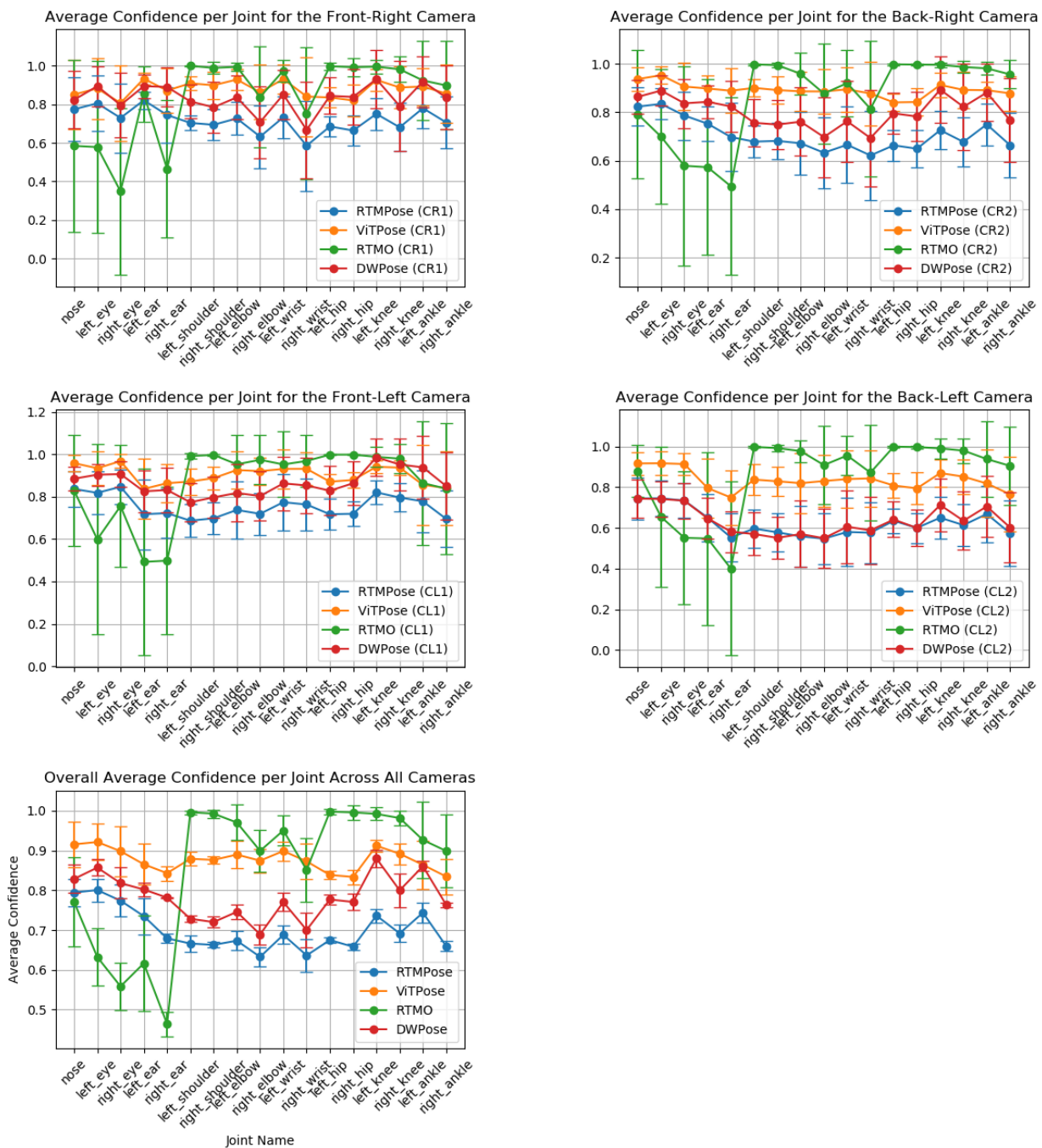




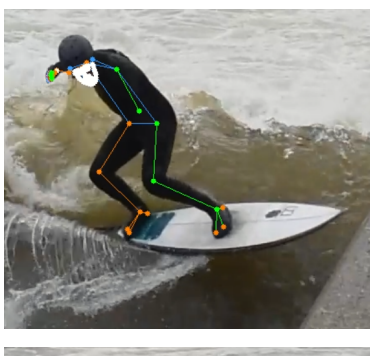
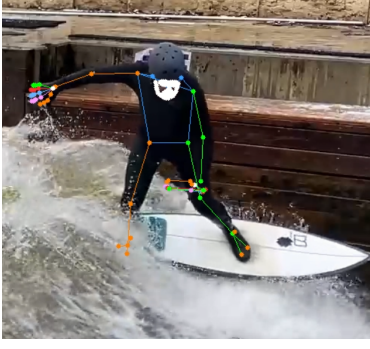



Figure 4.1: Composite figure showing the average confidence scores and standard deviations for the 2D pose estimation models across four camera views.

Table 4.2: Qualitative comparison of 2D pose estimation models under challenging scenarios

Model Name	Occlusion	Low Contrast
ViTPose [Xu22]		
RTMO [Lu23]		
DWPose [Yan23]		
RTMPose [Jia23]		

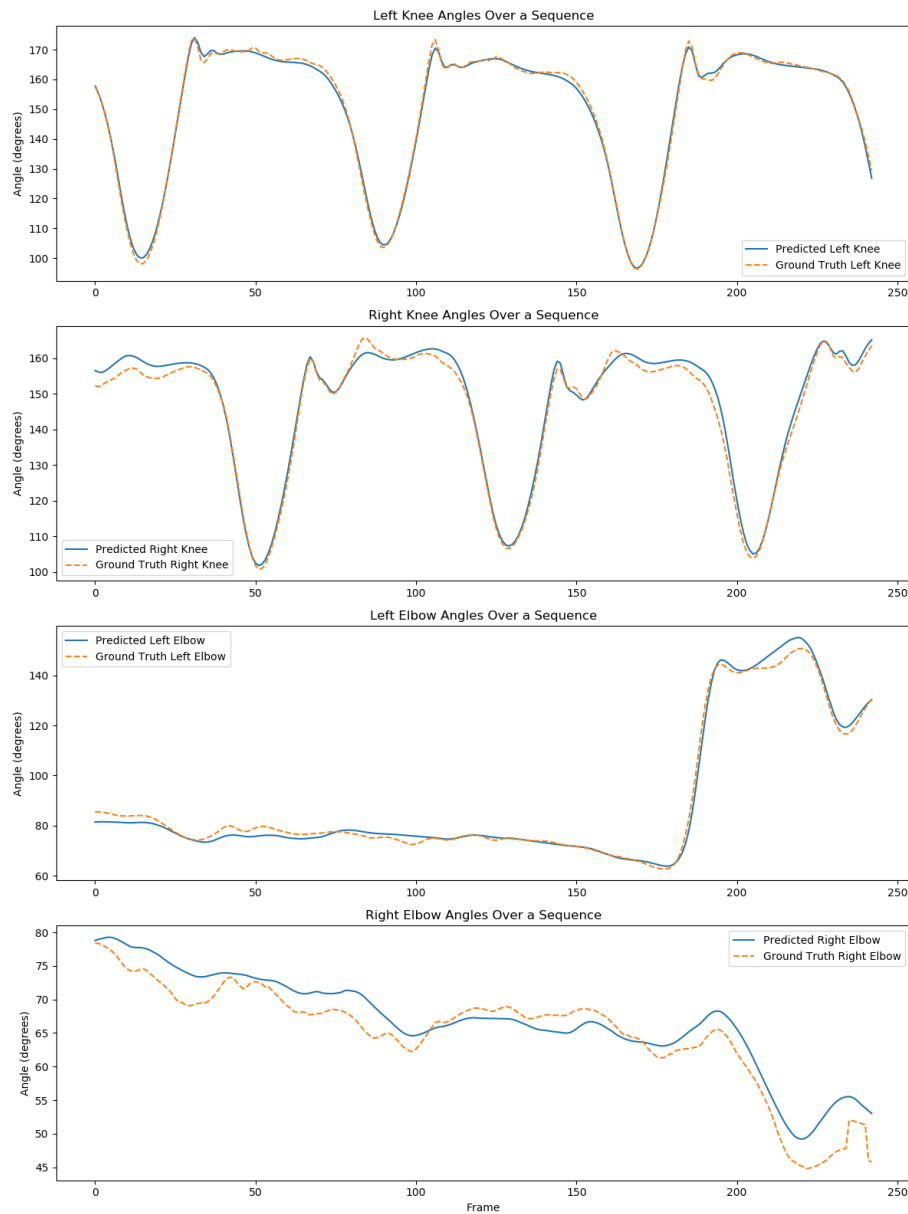


Figure 4.2: Joint angle comparison over a sequence of frames for left and right elbows and knees during a walking movement from the Human3.6M dataset, showing the MotionBERT model’s predictions closely align with ground truth measurements for the right and left knees and the left elbow. While the right elbow displays a higher discrepancy, the difference remains minor, within a few degrees, indicating overall good model performance.

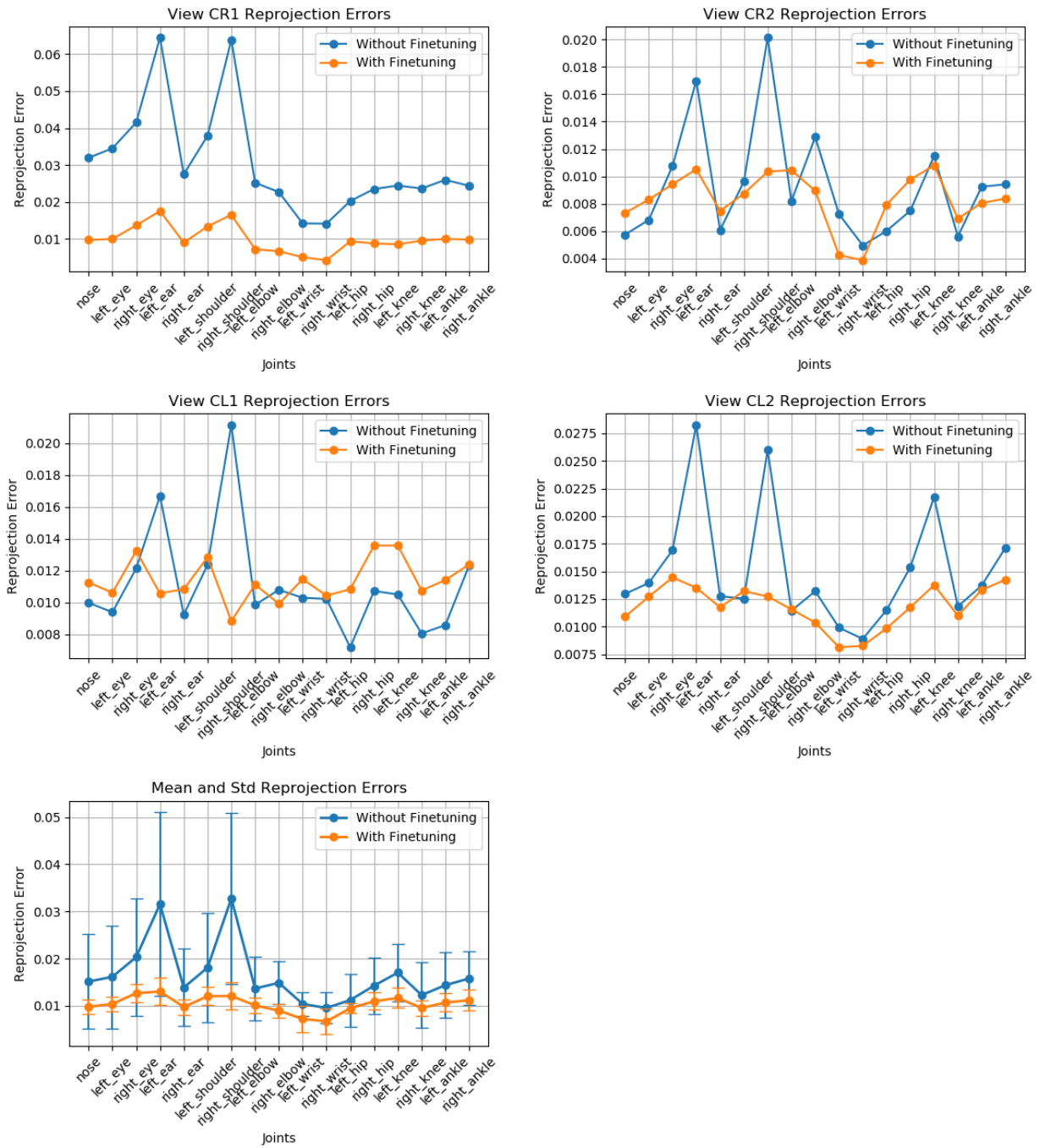


Figure 4.3: Comparison of reprojection errors across all camera views before and after model fine-tuning. The reprojection error is shown for each joint, indicating the improvements achieved through the fine-tuning process.

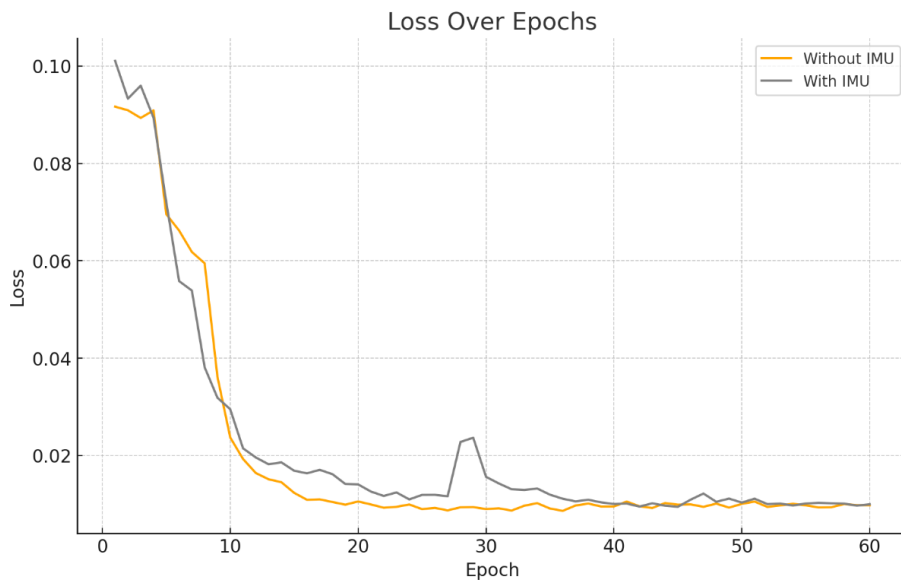


Figure 4.4: Testing set reprojection error of MotionBERT over 60 epochs with and without IMU data integration. The plot evaluates the hypothesis that IMU data can reduce reprojection error in 3D pose estimation.

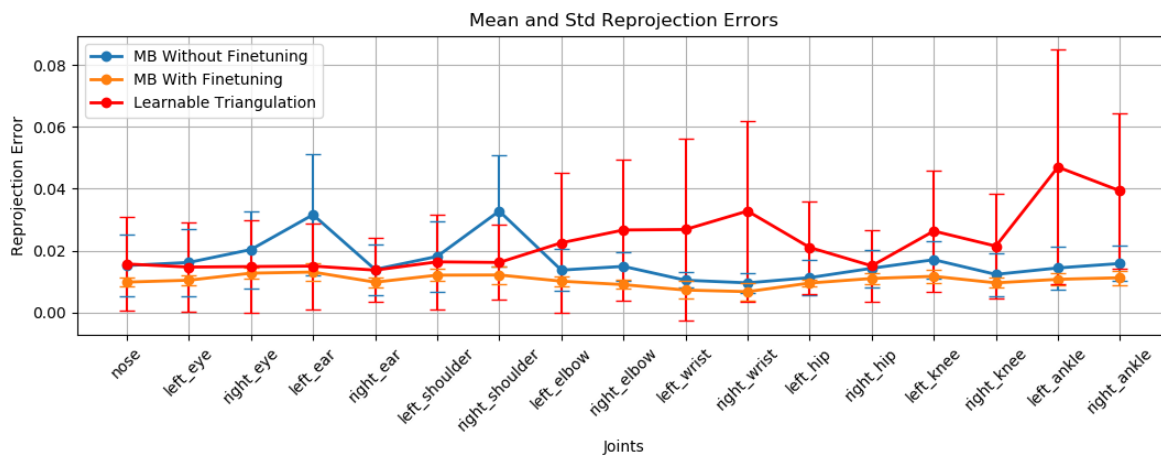


Figure 4.5: Comparison of reprojection errors using multi-view 3D pose estimation with the learnable triangulation model and monocular estimations from the MotionBERT and finetuned MotionBERT models.

Model	Average Error	Standard Deviation
MotionBERT (No Finetuning)	0.0166	0.0085
MotionBERT (With Finetuning)	0.0104	0.0020
Learnable Triangulation	0.0227	0.0192
MotionBERT (No Finetuning, Non-Face)	0.0154	0.0070
MotionBERT (With Finetuning, Non-Face)	0.0101	0.0020
Learnable Triangulation (Non-Face)	0.0260	0.0215

Table 4.3: Comparison of Reprojection Errors Across Models and Joints

Chapter 5

Discussion

The results of this study represent a significant step forward in the field of human pose estimation, particularly within the context of dynamic and complex environments such as surfing. By leveraging a combination of 2D and 3D pose estimation techniques, alongside the integration of IMU data, this research has uncovered new insights into the capabilities and limitations of current pose estimation models. This discussion aims to contextualize these findings within the broader implications for athletic training, sports performance enhancement, and potential real-world applications.

The findings from the 2D pose estimation analysis reveal a high degree of accuracy and reliability across the tested models, with ViTPose and RTMO showing particularly strong performance. However, the variance observed in confidence scores for certain joints, especially facial features, underlines the challenges posed by occlusions and environmental factors unique to water sports. The quantitative and qualitative assessments underscore the critical need for robust models capable of handling complex scenarios, such as wave occlusion and low-contrast conditions.

These insights highlight the potential for these models to contribute to athletic training protocols by providing accurate feedback on an athlete's posture and technique in real-time. Moreover, the ability to accurately track and analyze poses under challenging conditions opens up new avenues for research and application in other sports and activities where environmental factors play a significant role.

The advancements in 3D pose estimation demonstrated through this study, particularly with the finetuned MotionBERT model, mark a considerable improvement in the field. The reduction in reprojection errors signifies not only an enhancement in the model's accuracy but also its practical applicability in analyzing complex movements. The success of the finetuning process, which integrated data from the Human3.6M dataset and the surf dataset, illustrates the value of

incorporating diverse datasets to improve model generalization.

The potential applications of these findings extend beyond sports analytics, offering implications for injury prevention and rehabilitation. By providing precise 3D models of athletes' movements, coaches and health professionals can identify and correct technique flaws that may lead to injury, thereby enhancing athlete safety and performance longevity.

The investigation into the integration of IMU data with pose estimation models did not yield the anticipated improvement in reprojection error. One plausible explanation for this result is the architecture used for incorporating IMU data into the pose estimation process. The current framework might not have been optimized to fully exploit the rich motion context provided by IMUs, suggesting a need for architectures specifically designed to leverage such multimodal data effectively. Additionally, the efficacy of IMU integration could be contingent on the availability of true 3D pose data from motion capture systems during the training phase. Without the foundational training on accurate 3D poses, the model may struggle to translate the additional context provided by IMUs into improved accuracy for complex activities like surfing. This insight points to a significant area for future research: developing and testing new model architectures or training approaches that can better harness IMU data alongside visual inputs.

Furthermore, the comparison between monocular (uplifting) and multi-view 3D pose estimation approaches revealed a distinct advantage of the former, particularly demonstrated by the finetuned MotionBERT model. A key factor contributing to this outcome is MotionBERT's utilization of temporal information, allowing for a more nuanced understanding and prediction of human movement over time. This capability is critical in sports scenarios characterized by rapid and complex motions, where temporal consistency can significantly enhance pose estimation accuracy. Additionally, the observed performance disparity may also stem from sub-optimal camera calibration parameters in the multi-view setup. Inaccurate calibration can lead to errors in estimating the spatial relationship between cameras and the scene, impacting the model's ability to triangulate 3D poses accurately.

This research elucidates the complexities and challenges of pose estimation in dynamic environments, offering valuable insights and advancements in the field. The findings underscore the importance of model robustness, the potential of fine-tuning with diverse datasets, and the nuanced role of multimodal data integration in improving pose estimation accuracy.

Chapter 6

Conclusion

In this thesis, we embarked on an exploratory journey to enhance the domain of sports analytics, with a specific focus on surfing, through the lens of 3D pose estimation. The synthesis of literature reviews, methodological rigor, and empirical analyses has culminated in a comprehensive understanding of the challenges and opportunities within this niche yet burgeoning field of study. Our research has not only illuminated the complexities inherent in capturing and analyzing the dynamic maneuvers of surfing but also highlighted the transformative potential of integrating advanced pose estimation technologies in sports training and performance evaluation.

The development and validation of a novel pose estimation framework, tailored to the dynamic and fluid environment of surfing, marks a significant milestone in the intersection of sports science and artificial intelligence. By leveraging multi-view camera synchronization, IMU data integration, and cutting-edge machine learning models, this study has demonstrated the feasibility of achieving high-fidelity motion capture in outdoor sports settings, a task that has traditionally been fraught with challenges such as occlusion, variable lighting conditions, and the unpredictable nature of water sports.

The quantitative and qualitative analyses presented in the results section reveal a landscape where accuracy and adaptability intersect, highlighting the potential of advanced computational models to revolutionize athletic training, performance enhancement, and injury prevention. The study's nuanced approach to 2D and 3D pose estimation underscores the importance of precision and reliability, demonstrating the significant strides made in tackling the complexities of pose estimation in water sports.

One of the key takeaways from this thesis is the critical role of fine-tuning and model adaptation in improving pose estimation accuracy. The success of the finetuned MotionBERT model, in particular, exemplifies how targeted adjustments and the integration of diverse datasets can lead

to improvements in model performance. This approach not only enhances the model's applicability across various conditions but also paves the way for its adoption in a wide range of real-world applications.

Despite the promising advancements, the exploration into the integration of IMU data presents a more complex picture. While the anticipated improvements in pose estimation accuracy were not realized, this inquiry opens up new avenues for research into the integration of multimodal data sources. The lessons learned here emphasize the importance of continued innovation and experimentation in the quest for more sophisticated and effective pose estimation models.

However, the journey does not end here. The horizon of sports analytics and pose estimation is ever-expanding, with numerous avenues for future research. Potential directions include the exploration of deep learning models capable of real-time pose estimation, the integration of environmental variables such as wave dynamics in the analysis of surfing performance, and the extension of this framework to other water-based sports. Furthermore, the ethical considerations surrounding the use of such technologies in sports, particularly in terms of privacy and data security, warrant thorough investigation and discussion.

In conclusion, this thesis represents a foundational step towards the realization of advanced tracking and feedback systems in sports. Through the lens of surfing, we have demonstrated the utility and potential of 3D pose estimation technologies in enhancing athletic performance, paving the way for future innovations in sports analytics. As we stand on the precipice of this technological frontier, it is our hope that this research will inspire further exploration and development within this exciting field.

List of Figures

1.1	1897 depiction from Captain Cook’s Voyages, illustrating early Hawaiian surfing, prior to the sport’s 19th-century suppression and later resurgence. [Coo97]	2
2.1	Diagram displaying frameworks for 2D Human Pose Estimation involving multiple persons	7
2.2	Overview of the Multi-Stage CNN for Pose Estimation	8
2.3	The RTMPose Architecture Overview	9
2.4	TPD Pipeline of DWPose	10
2.5	Overview of the proposed D3DP method	12
2.6	Architecture of the Model: DSTformer features dual-stream-fusion modules . . .	13
2.7	Architecture and components of MotionAGFormer	15
2.8	Method using algebraic triangulation and learned confidence levels for pose estimation	16
2.9	Conceptual framework of the volumetric triangulation method for pose estimation	17
2.10	Schematic representation of the proposed 3D human mesh reconstruction method	18
2.11	Overview of the Pose Estimation Pipeline using SMPL model and IMUs	19
3.1	Measurement of surfboard dimensions	27
3.2	Detailed View of the Camera and Smartphone Setup for Surfing Documentation .	29
3.9	The proposed strategy for joint training on the Human3.6M and surf datasets, focusing on fine-tuning the MotionBERT model	36
3.10	Illustration of integrating IMU data with 2D pose inputs	38
3.3	Satellite view of the camera setup	41
3.4	Synchronized multi-view camera views of a surfer	42
3.5	Checkerboard corners detection from the back-left and back-right camera views .	42
3.6	Overlay of the animated IMU data plot on the video footage, illustrating the synchronization process	43

3.7	Illustration of a failure case of AlphaPose	43
3.8	Comparative Analysis of 3D Pose Estimation Models on the Human3.6M Dataset	44
4.1	Composite figure showing the average confidence scores and standard deviations for the 2D pose estimation models across four camera views.	53
4.2	Joint angle comparison over a sequence of frames for left and right elbows and knees	55
4.3	Comparison of reprojection errors across all camera views	56
4.4	Testing set reprojection error of MotionBERT over 60 epochs	57
4.5	Comparison of reprojection errors using multi-view 3D pose estimation	57

List of Tables

2.1	Comparison of 3D Pose Estimation Models	20
3.1	IMUs attachment position table of the protocol	27
3.2	Protocol calibration Movements for IMUs	28
4.1	Overview of average confidences (\pm standard deviation) across all joints and non-face joints	46
4.2	Qualitative comparison of 2D pose estimation models under challenging scenarios	54
4.3	Comparison of Reprojection Errors Across Models and Joints	58

Bibliography

- [24a] *COCO-WholeBody Benchmark (2D Human Pose Estimation)*. Accessed: 2024-02-18. 2024. URL: <https://paperswithcode.com/sota/2d-human-pose-estimation-on-coco-wholebody-1>.
- [24b] *Fuchslochwelle Nürnberg*. Accessed: 2024-02-18. 2024. URL: <https://www.nuernberger-dauerwelle.de/fuchslochwelle/>.
- [Bad21] Aritz Badiola-Bengoa and Amaia Mendez-Zorrilla. “A Systematic Review of the Application of Camera-Based Human Pose Estimation in the Field of Sport and Physical Exercise”. In: *Sensors* 21.18 (2021). ISSN: 1424-8220. DOI: 10.3390/s21185996. URL: <https://www.mdpi.com/1424-8220/21/18/5996>.
- [Bor18] Marcio Borgonovo-Santos. “SURF BIOMECHANICS AND BIOENERGETICS”. PhD thesis. Dec. 2018. DOI: 10.13140/RG.2.2.36260.30088.
- [Bri19] Lewis Bridgeman, Marco Volino, Jean-Yves Guillemaut, and Adrian Hilton. “Multi-Person 3D Pose Estimation and Tracking in Sports”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2019.
- [Bur13] Magnus Burenius, Josephine Sullivan, and Stefan Carlsson. “3D Pictorial Structures for Multiple View Articulated Pose Estimation”. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 3618–3625. DOI: 10.1109/CVPR.2013.464.
- [Cao17] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. “Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”. In: *CVPR*. 2017.
- [Chu23a] Sungho Chun, Sungbum Park, and Ju Yong Chang. *Representation learning of vertex heatmaps for 3D human mesh reconstruction from multi-view images*. 2023. arXiv: 2306.16615 [cs.CV].

- [Chu23b] Sungho Chun, Sungbum Park, and Ju Yong Chang. *Representation learning of vertex heatmaps for 3D human mesh reconstruction from multi-view images*. 2023. arXiv: 2306.16615 [cs.CV].
- [Con20] MMPose Contributors. *OpenMMLab Pose Estimation Toolbox and Benchmark*. <https://github.com/open-mmlab/mmpose>. 2020.
- [Coo97] Captain James Cook. *Illustration: Surf-Swimming, Sandwich Islands (Hawaii)*. Accessed: date. 1897. URL: <https://healthandfitnesshistory.com/images/illustration-surf-swimming-sandwich-islands-hawaii/>.
- [Dep24] Statista Research Department. *Surfing - statistics & facts*. Accessed: 2024-02-10. 2024. URL: <https://www.statista.com/topics/9833/surfing/#topicOverview>.
- [Eic12] Marcin Eichner, Manuel Marin-Jimenez, Andrew Zisserman, and Vittorio Ferrari. “2D articulated human pose estimation and retrieval in (almost) unconstrained still images”. In: *IJCV* 99 (2012), pp. 190–214.
- [Fan22] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. *AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time*. 2022. arXiv: 2211.03375 [cs.CV].
- [Far17] Oliver RL Farley, Chris R Abbiss, and Jeremy M Sheppard. “Performance analysis of surfing: a review”. In: *The Journal of Strength & Conditioning Research* 31.1 (2017), pp. 260–271.
- [Fer94] Franco Ferraris, I Gorini, U Grimaldi, and Marco Parvis. “Calibration of three-axial rate gyros without angular velocity standards”. In: *Sensors and Actuators A: Physical* 42.1-3 (1994), pp. 446–449.
- [He20] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. *Epipolar Transformers*. 2020. arXiv: 2005.04551 [cs.CV].
- [Hua20a] Fuyang Huang, Ailing Zeng, Minhao Liu, Qiuxia Lai, and Qiang Xu. “DeepFuse: An IMU-Aware Network for Real-Time 3D Human Pose Estimation from Multi-View Image”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Mar. 2020.

- [Hua20b] Junjie Huang, Zheng Zhu, Feng Guo, and Guan Huang. “The devil is in the details: Delving into unbiased data processing for human pose estimation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2020.
- [Ion14] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. “Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7 (July 2014), pp. 1325–1339.
- [Isk19] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. *Learnable Triangulation of Human Pose*. 2019. arXiv: 1905.05754 [cs.CV].
- [Jia23] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. *RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose*. 2023. arXiv: 2303.07399 [cs.CV].
- [Jin20] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. “Whole-Body Human Pose Estimation in the Wild”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2020.
- [Kau24] Satya Mallick Kaustubh Sadekar. *Camera calibration using opencv*. Jan. 2024. URL: <https://learnopencv.com/camera-calibration-using-opencv/>.
- [Li22] Yanjie Li, Sen Yang, Peidong Liu, Shoukui Zhang, Yunxiao Wang, Zhicheng Wang, Wankou Yang, and Shu-Tao Xia. *SimCC: a Simple Coordinate Classification Perspective for Human Pose Estimation*. 2022. arXiv: 2107.03332 [cs.CV].
- [Lin14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. “Microsoft coco: Common objects in context”. In: *ECCV*. 2014.
- [Lin23] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. “One-Stage 3D Whole-Body Mesh Recovery with Component Aware Transformer”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.
- [Lop15] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. “SMPL: A Skinned Multi-Person Linear Model”. In: *ACM Transactions on Graphics (TOG)* 34.6 (2015), pp. 1–16.

- [Lu23] Peng Lu, Tao Jiang, Yining Li, Xiangtai Li, Kai Chen, and Wenming Yang. *RTMO: Towards High-Performance One-Stage Real-Time Multi-Person Pose Estimation*. 2023. arXiv: 2312.07526 [cs.CV].
- [Lyu22] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. *RTMDet: An Empirical Study of Designing Real-Time Object Detectors*. 2022. DOI: 10.48550/ARXIV.2212.07784. arXiv: 2212.07784 [cs.CV]. URL: <https://arxiv.org/abs/2212.07784>.
- [Meh23] Soroush Mehraban, Vida Adeli, and Babak Taati. *MotionAGFormer: Enhancing 3D Human Pose Estimation with a Transformer-GCNFormer Network*. 2023. arXiv: 2310.16288 [cs.CV].
- [Nen17] Jim Nendel. “Surfing in early twentieth-century Hawai’i: The appropriation of a transcendent experience to competitive American sport”. In: *Sport in the Pacific*. Routledge, 2017, pp. 122–136.
- [Pav18] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. “Learning to Estimate 3D Human Pose and Shape from a Single Color Image”. In: *CVPR*. 2018.
- [Red21] N Dinesh Reddy, Laurent Guigues, Leonid Pishchulin, Jayan Eledath, and Srinivasa G. Narasimhan. “TesseTrack: End-to-End Learnable Multi-Person Articulated 3D Pose Tracking”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 15190–15200.
- [Rho18] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. “Learning Monocular 3D Human Pose Estimation From Multi-View Images”. In: *CVPR*. 2018.
- [Sha23] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. *Diffusion-Based 3D Human Pose Estimation with Multi-Hypothesis Aggregation*. 2023. arXiv: 2303.11579 [cs.CV].
- [Tos14] Alexander Toshev and Christian Szegedy. “Deeppose: Human pose estimation via deep neural networks”. In: *CVPR*. 2014.
- [Von18] 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)*. 2018, pp. 601–617.

- [Wan19] Jianbo Wang, Kai Qiu, Houwen Peng, Jianlong Fu, and Jianke Zhu. “AI Coach: Deep Human Pose Estimation and Analysis for Personalized Athletic Training Assistance”. In: *Proceedings of the ACM International Conference on Multimedia (ACM MM)*. 2019.
- [Wen19] Chuang Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. “Photo Wake-Up: 3D Character Animation From a Single Photo”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.
- [Wen23] Christina Wende, Christophe Lambert, Juergen Hoehner, and Maurice Balke. “Injuries and use of safety equipment in river surfing”. In: *Orthopaedic journal of sports medicine* 11.4 (2023), p. 23259671231155884.
- [Xu22] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. *ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation*. 2022. arXiv: 2204.12484 [cs.CV].
- [Yan12] Yi Yang and Deva Ramanan. “Articulated human detection with flexible mixtures of parts”. In: *IEEE TPAMI* 35 (2012), pp. 2878–2890.
- [Yan23] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. *Effective Whole-body Pose Estimation with Two-stages Distillation*. 2023. arXiv: 2307.15880 [cs.CV].
- [Zha20] Zhe Zhang, Chunyu Wang, Weichao Qiu, Wenhui Qin, and Wenjun Zeng. “AdaFuse: Adaptive Multiview Fusion for Accurate Human Pose Estimation in the Wild”. In: *International Journal of Computer Vision* 129.3 (Nov. 2020), pp. 703–718. ISSN: 1573-1405. DOI: 10.1007/s11263-020-01398-9. URL: <http://dx.doi.org/10.1007/s11263-020-01398-9>.
- [Zhe20] Ce Zheng, Wenhan Wu, Taojiannan Yang, Sijie Zhu, Chen Chen, Ruixu Liu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. “Deep Learning-Based Human Pose Estimation: A Survey”. In: *CoRR* abs/2012.13392 (2020). arXiv: 2012.13392. URL: <https://arxiv.org/abs/2012.13392>.
- [Zhu23] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. *MotionBERT: A Unified Perspective on Learning Human Motion Representations*. 2023. arXiv: 2210.06551 [cs.CV].
- [Zöl23] Michael Zöllner, Moritz Krause, Jan Gemeinhardt, Michael Döllinger, and Stefan Kniesburges. “Evaluation of Machine Learning based Pose Estimation of Surfers on River Waves”. In: *Proceedings of the 16th International Conference on Pervasive Technologies Related to Assistive Environments. PETRA '23.*, Corfu, Greece, Asso-

ciation for Computing Machinery, 2023, pp. 443–447. ISBN: 9798400700699. DOI: 10.1145/3594806.3596570. URL: <https://doi.org/10.1145/3594806.3596570>.

Appendix A

Acronyms

DNN deep neural network

HPE human pose estimation

2D two-dimensional

3D three-dimensional

PAF part affinity field

TPD two-stage pose distillation

KD knowledge distillation

MSE mean squared error

GAU gated attention unit

CNN convolutional neural network

IMU inertial measurement unit

D3DP diffusion-based 3D pose estimation

JPMA joint-wise reprojection-based multi-hypothesis aggregation

MHSA multi-head self-attention

GCN graph convolutional network

MLP multi-layer perceptron

DSTformer dual-stream spatio-temporal transformer

AP average precision

VHA vertex heatmap autoencoder

BCP body code predictor

SMPL skinned multi-person linear model

LMT learnable human mesh triangulation

MPJPE mean per joint position error

mAP mean Average Precision

MoCap motion capture