





Component Replacement Study of 3D Human Pose Estimation Models in Real-World Complex Sports Scenarios: Focusing on Head Impact Events

Yuchen Shi¹^a, Nobutake Ozeki²^b, Ryu Yoshida², Motoki Inaji³^c,
Kazuyoshi Yagishita⁴ and Yusuke Miyazaki¹^d

¹Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology,
2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan

²Department of Orthopaedic Surgery, Tokyo Medical and Dental University,
1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan

³Department of Functional Neurosurgery, Tokyo Medical and Dental University,
1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan

⁴Clinical Center for Sports Medicine and Sports Dentistry, Tokyo Medical and Dental University,
1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan

Keywords: Human Pose Estimation, Head Impact Events, Complex Sports Scenarios, Concussion Biomechanics, Monocular Camera, Deep Learning, Computer Vision.


Abstract: Human pose estimation in 3D is crucial in complex sports scenarios, particularly for athlete head impact events. We investigated the effect of different 2D pose estimation methods on the performance of 3D pose estimation models in complex sports environments. We used a transformer-based 3D human pose estimation model as a base framework, creating multiple variants by replacing the 2D pose estimator. These variants were evaluated using real sports game videos. Four 2D pose estimators were employed: Simple Baseline, High-Resolution Network (HRNet), Multi-stage Pose Network (MSPN), and Residual Steps Network (RSN). Performance was assessed using Mean per Joint Positional Error (MPJPE), Procrustes analysis MPJPE (P-MPJPE), and Mean per Joint Velocity Error (MPJVE) metrics. The results showed that MSPN performed the best in terms of position accuracy and motion velocity consistency (MPJPE, P-MPJPE and MPJVE). RSN presented promising absolute position accuracy (MPJPE) but showed limitations in the overall pose configuration (P-MPJPE). Simple Baseline and HRNet proved to be inadequate for complex sports scenarios. These findings indicate that different model architectures have different advantages in 3D human pose estimation in complex sports scenarios. This study provides insights for improving 3D pose estimation models in challenging real-world sports applications, contributing to the better understanding and prevention of sports-related head injuries.


1 INTRODUCTION


1.1 Background


Concussion in sport is a critical issue in modern sports medicine due to its severe impact on athletes' health

and careers. Annually, an estimated 4 million sport-induced concussions occur from rapid brain impacts (Bryan et al., 2016). These injuries cause cognitive impairment and, functional brain changes, and increase the risk of further injury (Giza & Hovda, 2001; McKee et al., 2013; Courtney & Courtney,

^a <https://orcid.org/0000-0001-8568-5229>

^b <https://orcid.org/0000-0002-2927-7930>

^c <https://orcid.org/0000-0002-6759-5729>

^d <https://orcid.org/0000-0001-7863-7704>

2015). However, the precise biomechanical mechanisms of sports-related concussions remain unclear, hindering the development of effective prevention strategies (Ji et al., 2015).

To mitigate concussion risks in sports, understanding head impact dynamics is essential (Camarillo et al., 2013). Finite element (FE) simulations have become crucial in biomechanical analyses, elucidating the mechanical forces and brain tissue deformation in concussive events (Madhukar & Ostoja-Starzewski, 2019). Accurate kinematic inputs, such as impact velocity and location, are critical for realistic simulations and a better understanding of concussion causes and effects. Consequently, precisely capturing head motions during impacts for FE simulations has become a key research focus.

Traditional head impact kinematics measurements use optical markers or sensors attached to athletes (Camarillo et al., 2013; Cortes et al., 2017). However, these methods are invasive, interfere with natural movements, and are impractical in real-world sports settings (King et al., 2015; Wu et al., 2016). To overcome these limitations, non-contact measurement techniques using monocular 2D video data are becoming necessary. Such methods would enable practical and effective impact measurements without burdening athletes or requiring extensive camera equipment.

Quantifying head impact kinematics from 2D monocular video involves two main phases: 1) either a multi-stage process (video acquisition, 2D pose estimation, and 2D-to-3D upgrade) or a single-stage approach (direct 3D pose estimation from video); and 2) reconstructs of 3D human motion and determination of head impact kinematics based on the 3D pose or shape predicted in the first phase.

We focused on the multi-stage process of quantifying head impact kinematics, specifically extracting 2D human poses and lifting them to 3D. Most high-performing 3D human pose estimation methods use this framework, relying heavily on 2D pose estimation techniques (Moon et al., 2019; Rogez et al., 2020; Liu et al., 2022). Different 2D pose estimation methods significantly affect the overall 3D pose estimation performance. By quantifying these performance differences and analyzing their effects, we aimed to provide an objective basis for method selection and optimization in constructing 3D pose estimation models.

Previous computer vision research has developed advanced deep learning models for 2D and 3D pose estimation (Newell et al., 2016; Cao et al., 2017; Pavlakos et al., 2017; Xiao et al., 2018; Sun et al., 2019; Li et al., 2022). These supervised learning

models are typically trained and tested on standard dataset videos before their application in realistic scenes. However, significant differences exist between standard datasets and actual sports videos, affecting model performance in real-world scenarios:

1. Video quality and consistency: Real games are affected by weather, filming techniques, and lighting, unlike controlled standard datasets.

2. Camera angles: Actual games are filmed from multiple angles, while standard datasets use optimal or fixed viewpoints.

3. Scene complexity: Real games involve spectator interference and multiple simultaneous plays, contrasting with the simpler, controllable standard dataset videos.

4. Data diversity: Real game videos offer genuine diversity but may suffer from insufficient data collection, while standard datasets simulate diversity but may have inherent selection biases.

These differences underscore the importance of evaluating and refining computer vision models in real-world applications. Assessing model performance on real scene videos provides a comprehensive understanding of real-world applicability, forming a crucial basis for model refinement and optimization.

1.2 Research Purpose

This study examined pose estimation in athlete head impact events, focusing on how different 2D pose estimation methods affect 3D pose estimation performance in complex real-world sports scenarios. We used a multi-stage 3D human pose estimation model as a base framework, creating variants by replacing the top-down 2D pose estimator. These variants were evaluated on real sports scene videos. By analyzing the effects of different 2D methods on the overall 3D performance, we proposed strategies to improve 3D pose estimation, addressing challenges like fast movements, occlusions, and complex poses. We aimed to contribute to the development of robust and efficient 3D human pose estimation algorithms for complex real-world sports scenarios.

2 RELATED WORK

2.1 Single-Person 2D Human Pose Estimation

Single-person 2D pose estimation models typically employ regression-based (Toshev & Szegedy, 2014; Carreira et al., 2016) or detection-based approaches

(Newell et al., 2016; Wei et al., 2016). These frameworks generally consist of a pose encoder, which extracts high-level features from high to low resolution, and a pose decoder, which estimates 2D keypoints. Regression-based decoders directly output keypoint coordinates but struggle with complex poses due to non-linearity. Detection-based decoders generate keypoint heatmaps and are more robust in handling complex poses (Liu et al., 2022).

2.2 Multi-Person 2D Human Pose Estimation

Multi-person 2D pose estimation methods use either top-down (Xiao et al., 2018; Sun et al., 2019; Li et al., 2019; Cai et al., 2020) or bottom-up approaches (Cao et al., 2017; Cheng et al., 2020). Top-down methods first localize individuals, then apply single-person pose estimation to each person. Bottom-up methods predict all keypoints simultaneously, then assign them to individuals. In videos, top-down approaches detect and predict keypoints frame-by-frame, propagating them across frames. Bottom-up methods predict all keypoints per frame, then assign them to individuals using spatio-temporal patterns.

The top-down multi-person 2D pose estimation approach has several advantages. It could utilize a specialized single-person pose estimation technique that focuses on only one person at a time within the detected bounding box, thus achieving highly accurate keypoint localization for a single person. This method isolates each person and reduces background interference and is therefore robust to cluttered backgrounds. In addition, this approach could be integrated with existing advanced object detection frameworks (Faster R-CNN [Ren et al., 2015] or YOLO [Redmon et al., 2016]) to take full advantage of their benefits. The segmented processing pipeline (detection followed by pose estimation) also facilitates individual optimization of each module, thus improving the overall performance of the model.

3 METHOD

3.1 Multi-Stage Approach for 3D Human Pose Estimation

We aimed to examine how different 2D pose estimation methods affect 3D model performance in complex real sports scenarios. We used a multi-stage 3D human pose estimation model as a base

framework, creating multiple variants by replacing the 2D pose estimators within the model.

This multi-stage 3D human pose estimation model consisted of two main stages, as illustrated in Figure 1. In the first stage, a multi-person 2D human pose extractor processed monocular video frames to extract 2D pose sequences. The second stage then took these 2D pose sequences as input and employed a 3D human pose estimation model to reconstruct corresponding 3D human poses.

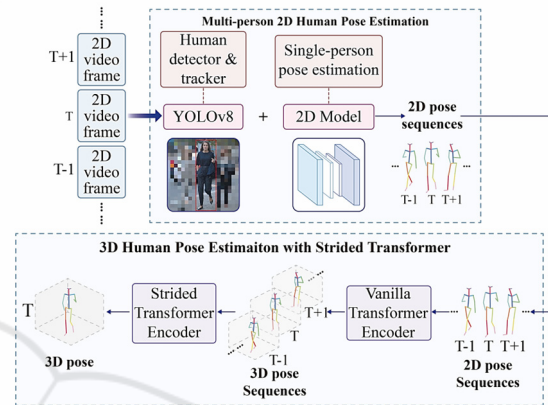


Figure 1: The proposed multi-person 3D human pose estimation framework.

The multi-person 2D human pose detection stage in our model employed a top-down approach comprising two main tasks. First, for person detection and tracking, we utilized YOLOv8 (Jocher et al., 2023) to detect individuals in video frames, followed by the BoT-SORT algorithm (Aharon et al., 2022) to track each detected person frame-by-frame. Second, for single-person pose estimation, we cropped the area around each detected person based on their bounding box. These cropped frames were then input into a single-person 2D human pose estimation model, which estimated the pose of each individual.

In the 3D human pose estimation stage, we employed a transformer-based model (Li et al., 2022) to lift 2D pose sequences to 3D. This process involved two main components: the Vanilla Transformer Encoder (VTE) and the Strided Transformer Encoder (STE). The VTE processed the input 2D pose sequence, predicting the 3D pose sequence and capturing temporal information to ensure motion consistency. The STE then received the VTE output, utilizing strided convolutional layers instead of fully-connected layers in its feed-forward network. This architecture shortened the sequence length and effectively combined global context from self-attention with local context from strided convolution. Ultimately, the STE predicted the 3D

pose of the center frame and recovered the entire 3D pose sequence from contextual information.

We evaluated four detection-based top-down 2D pose estimation models as alternatives for the 3D pose estimation component: Simple Baseline (Xiao et al., 2018), High-Resolution Network (HRNet) (Sun et al., 2019), Multi-Stage Pose Network (MSPN) (Cai et al., 2020) and Residual Steps Network (RSN) (Li et al., 2019). While these models perform well on standard datasets, their effectiveness in complex sports scenarios, particularly athlete head impacts, remains unexplored. We focused on their architectural features to assess their potential when integrated into a 3D pose estimation model. The models used weights pre-trained on the COCO val2017 dataset. The transformer-based 2D-to-3D lifting models were pre-trained on Human3.6M and HumanEva-I datasets (Li et al., 2022).

3.2 Simple Baseline

Simple Baseline is a 2D human pose estimation model based on convolutional neural networks (Figure 2). The model first uses a ResNet as the backbone network, and a complete pose encoding-decoding network is constructed by adding a small number of deconvolutional layers after the backbone network. The encoding part learns a high-level feature representation of the human body pose, and the decoding part up-samples the feature maps according to the input image size to generate a heatmap of the keypoints of the human body.

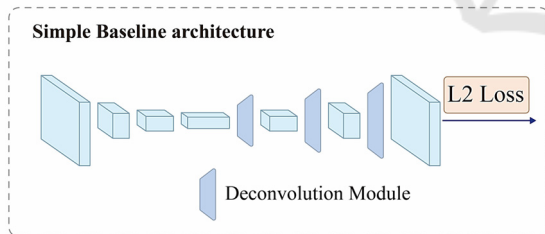


Figure 2: Simple Baseline architecture (Xiao et al., 2018).

3.3 HRNet

HRNet is a multi-stage, multi-branch fusion network architecture for 2D human pose estimation (Figure 3). Unlike traditional encoding-decoding architectures, HRNet maintains high-resolution feature representations throughout the network, avoiding the loss of spatial information due to down sampling. The network consists of multiple parallel sub-networks, each processing feature information at different levels. At each stage, multi-level features from different

branches are fused and interact with each other through cross-connections. As the network progresses, the high-resolution branch gradually integrates contextual information from the low-resolution branch while maintaining the high-resolution details required for keypoint localization. Through multi-level feature fusion and refinement, HRNet realizes the effective combination of global and local information. In the last stage of the network, the feature maps of all branches are summarized and up-sampled to the original resolution of the input image to generate a heatmap of keypoints.

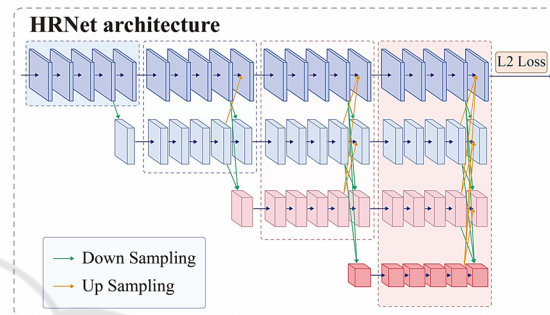


Figure 3: High-Resolution Network (HRNet) architecture (Sun et al., 2019).

3.4 MSPN

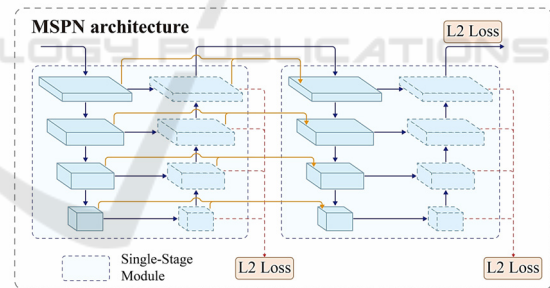


Figure 4: Multi-Stage Pose Network (MSPN) architecture (Cai et al., 2020).

MSPN is a multi-stage network structure for 2D human pose estimation (Figure 4). The network works by cascading multiple prediction stages, each of which receives the fusion information of the feature maps output from the previous stage and the feature maps at each level of the previous stage. Through this linking, each stage can optimize the prediction results from the previous stage, combining global and local information to refine keypoint locations. Through iterative optimization over multiple stages, the network can gradually improve the accuracy of keypoint localization. To better

instruct the network learning, MSPN introduces supervised signals at each intermediate stage, allowing the network to learn the multi-level feature representations required for the pose estimation task at different stages. In the final stage, the network generates a high-resolution heat map of the keypoints as the final output.

3.5 RSN

RSN is a multi-stage network architecture for 2D human pose estimation (Figure 5). The network cascades multiple residual steps to progressively refine and optimize the prediction of keypoints. Each RSN module in a residual step shares similarities with ResNet in its overall structure, but differs from ResNet in the structure of its component units. The RSN module consist of multiple Residual Steps Blocks (RSBs), which divide the input features into four branches, each with a different number of 3×3 convolutional layers (ranging from zero to three). Through the dense connected structure of these branches and 3×3 convolutional layers, the overall network has access to a wide range of receptive fields, making it well-qualified to learn delicate representations of the features, as well as capturing information about the features at a variety of different scales. In addition, RSN introduces supervised signals in the middle of each residual stage, allowing the network to learn meaningful feature representations at different stages.

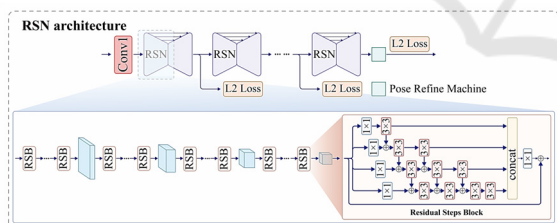


Figure 5: Residual Steps Network (RSN) architecture (Li et al., 2019).

After RSN undergoes feature fusion within the RSB and in each residual step, features at different levels are fused together, which contain both low-level precise local information and high-level global information. These features contribute differently to the final prediction results. To solve this problem, RSN proposes an efficient attention mechanism, the pose refine machine (PRM), to weight between the local and global representations of the output features to further refine the location of the keypoints, and ultimately output the keypoint heatmap.

4 EXPERIMENTS

4.1 Test Dataset

To evaluate the proposed human pose estimation variants in real sports scenarios, we used 15 rugby game video clips, that were approved by the ethics committees of Tokyo Medical and Dental University and Tokyo Institute of Technology. From these clips, we manually identified 16 significant impact events. Each event included 21 consecutive frames: 10 pre-impact, 1 impact, and 10 post-impact frames, totaling 336 impact-related frames, within which 218 target persons were detected and tracked. After data cleaning, our final multi-person pose dataset comprised 3,521 frames.

To create a reference standard for our study, we undertook a two-step process. First, we manually labeled 17 keypoints for each athlete in every frame, ensuring consistency with COCO dataset definitions. Subsequently, we lifted these 2D keypoint coordinates to 3D space using our transformer-based model, creating a comprehensive 3D representation of each pose. These manually labeled and 3D-lifted coordinates serve as the ground truth for our analysis.

4.2 Evaluation Metrics

We evaluated each 3D human pose estimation variant model using three key metrics: Mean per Joint Positional Error (MPJPE), Procrustes analysis MPJPE (P-MPJPE), and Mean per Joint Velocity Error (MPJVE). MPJPE (Ionescu et al., 2013) measures absolute positional accuracy by calculating the average Euclidean distance between ground truth and predicted joint positions. This metric can be used to evaluate the keypoint localization performance of models in the context of rapid movements by assessing the resilience to motion instability, complex postures, partial occlusions, and background interference. P-MPJPE (Martinez et al., 2017) provides a normalized accuracy assessment by aligning the estimated 3D pose with the ground truth before error calculation, allowing a fair comparison of pose estimates at different scales and orientations. This metric can be used to assess the proficiency of the model in capturing the overall pose structure and serves as a key indicator of its ability to interpret complex postures and adapt to diverse camera perspectives. MPJVE (Pavlo et al., 2019) employs the first-order derivative of MPJPE to assess the temporal smoothness of predicted results. This metric is particularly crucial for video-based quantification of head impact velocities, for which consistency in

motion estimation is of paramount importance. These metrics collectively provide a comprehensive evaluation of both positional accuracy and temporal consistency in real-world scenarios, especially for head impact events in complex sports settings.

4.3 Statistical Analysis

We conducted statistical analyses on the MPJPE, P-MPJPE, and MPJVE error samples for the four model variants to identify pairwise significant differences in performance. Our analysis process began with a Shapiro-Wilk test to check for normal distribution in all sample groups, followed by Levene's test to assess variance consistency between comparison groups. We then performed Kruskal-Wallis H-tests on the MPJPE, P-MPJPE, and MPJVE results for all four models. Statistical significance was set at p -value < 0.05 . This comprehensive analysis was used to thoroughly compare the performance of the four model variants across all three metrics.

4.4 Implementation Details

In our experiments, the Simple Baseline used ResNet as the backbone network with a depth of 152 layers. The initial number of channels of the HRNet model was set to 48. The MSPN model consisted of 4 cascaded single-stage modules; the overall network depth was 50 layers. The RSN model consisted of 3 residual steps in cascade; the overall network depth was 50 layers. Before using the 2D keypoints output from the four models as input to the 3D model, we converted the 2D keypoints from COCO format to H36M format. For 3D pose estimation, the transformer-based 3D human pose estimation model used 3 VTE and 3 STE encoder modules with the number of channels set to 256. The temporal motion kernel size and stride factor of the STE modules were set to 3. The receptive field of the model inputs was 27 frames, and the left and right padding operations were performed on inputs with < 27 frames to compensate for the complete number of frames.

5 RESULTS

In this study, we constructed four 3D human pose estimation models, each based on a different 2D pose estimation method. An example of the visual results of the four variant 3D human pose estimation models on our real-world rugby video dataset are presented in Figure 6. Table 1 summarizes the quantitative results of the three metrics for all four models.

The performance of the four variant models varied across the three metrics. For MPJPE, the MSPN-based model showed the lowest error (86.12 mm), closely followed by the RSN model (86.78 mm), with HRNet and Simple Baseline models showing slightly higher errors (87.62 mm and 87.55 mm, respectively). In terms of P-MPJPE, the MSPN model again outperformed the others with an error of 55.80 mm, while the remaining models had errors between 56.72 and 57.48 mm. For MPJVE, the MSPN model performed best (83.92 mm/frame), followed by RSN (85.27 mm/frame), with Simple Baseline and HRNet showing slightly higher velocity errors (85.68 mm/frame and 86.22 mm/frame, respectively).

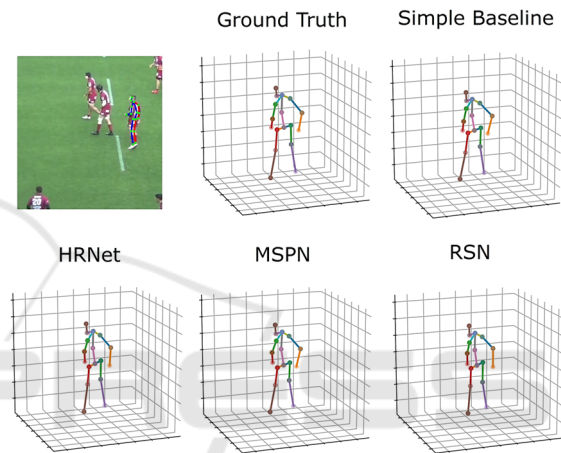


Figure 6: Visual results of four variant 3D human pose estimation models on real-world rugby video dataset.

To assess the statistical significance of the observed differences in metrics, we conducted a series of tests. The Shapiro-Wilk test revealed that none of the sample groups followed a normal distribution. Levene's test indicated inconsistent sample variances for all three metrics across the four groups. Consequently, we employed the Kruskal-Wallis H test followed by pairwise comparisons. The Kruskal-Wallis H test demonstrated significant differences among the four models for all three metrics ($p < 0.05$ for MPJPE, P-MPJPE and MPJVE), indicating that the performance variations between the models were statistically meaningful.

Real-world sports scenarios are characterized by dynamic variables such as fluctuating lighting conditions, intermittent occlusions, and rapid movement patterns. Such environmental diversity introduces significant sample heterogeneity, potentially yielding counterintuitive statistical results. In these cases, the larger mean differences lack significance whereas the smaller ones are significant.

To address this concern, our analysis integrated statistical significance with actual sample values, aiming for a balanced and accurate interpretation of the observed data.

As shown in Figure 7, pairwise comparisons resulted in the following conclusions:

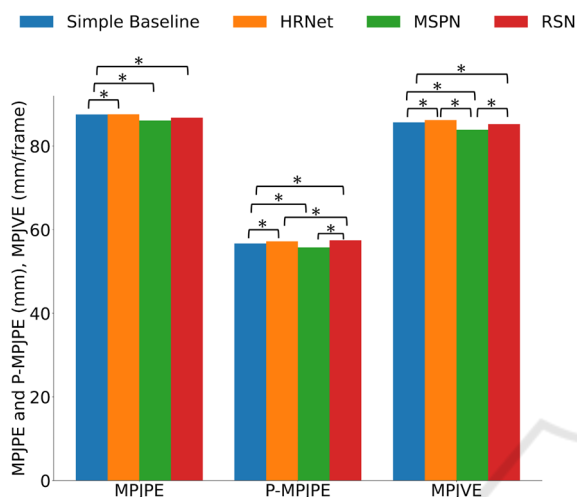


Figure 7: Statistical results of the four models on MPJPE, P-MPJPE and MPJVE. * p-value < 0.05.

- HRNet-based model underperformed Simple Baseline-based model in MPJPE, P-MPJPE and MPJVE (all $p < 0.05$).
- MSPN-based model outperformed Simple Baseline-based model in MPJPE, P-MPJPE and MPJVE (all $p < 0.05$).
- RSN-based model outperformed Simple Baseline-based model in MPJPE and MPJVE but underperformed in P-MPJPE (all $p < 0.05$).
- HRNet-based model underperformed MSPN-based model in MPJPE, P-MPJPE (although not statistically significant) and MPJVE ($p < 0.05$).
- HRNet-based model outperformed RSN-based model in P-MPJPE ($p < 0.05$) but underperformed in MPJPE and MPJVE (although not statistically significant).
- MSPN-based model outperformed RSN-based model in P-MPJPE and MPJVE ($p < 0.05$), but showed no significant differences in MPJPE ($p = 0.48$).

Table 1: MPJPE (mm), P-MPJPE (mm) and MPJVE (mm/frame) results for the four variant models.

	MPJPE	P-MPJPE	MPJVE
Simple Baseline-based	87.55	56.72	85.68
HRNet-based	87.62	57.24	86.22
MSPN-based	86.12	55.80	83.92
RSN-based	86.78	57.48	85.27

6 DISCUSSION

In this study, we employed a multi-stage 3D human pose estimation model as our base framework. We created several variant models by systematically replacing the top-down 2D pose estimator within this framework and evaluated their performance using real sports scenario videos. Our comparison focused on four 3D human pose estimation models, each utilizing a different 2D pose estimation methodology: Simple Baseline, HRNet, MSPN and RSN. Through this approach, we were able to assess the unique capabilities and limitations of these 2D pose estimation models when applied to the task of 3D pose estimation from 2D inputs in complex, real-world sports scenarios.

The MSPN-based and RSN-based variants presented superior performance in MPJPE (86.12 mm and 86.78 mm, respectively). The MSPN-based variants also performed excellently in terms of the MPJVE (83.92 mm) and P-MPJPE (55.80 mm) metrics. The exceptional performance of the MSPN and RSN models can be attributed to their multi-stage cascade structure. This architecture enabled iterative refinement of keypoint locations across multiple stages, which progressively enhanced localization accuracy. Moreover, a key feature of MSPN and RSN is the introduction of supervised signals at each intermediate stage, which compelled every level of the network to generate robust and reliable feature representations, rather than relying solely on the final output layer. Furthermore, the MSPN and RSN incorporate a sophisticated multi-level feature fusion mechanism at each cascade stage, which effectively integrated local and global features. Global features provided information about the overall body configuration to facilitate an understanding of the relative positions of different body segments. Concurrently, local features focused on specific keypoints or joint regions to offer precise localization details. In addition, the RSB in the RSN model employs a densely connected structure of branches and convolutional layers. This structure enabled the network to obtain a wide receptive field and generate fine feature representations, while simultaneously

capturing feature information at various scales. These structural characteristics significantly facilitated the model's feature representation capabilities, thereby enhancing both the accuracy and robustness of the model and allowing the model to locate keypoint positions more accurately when confronted with fast movements, complex postures and occlusion situations.

Despite the strong performance RSN in absolute keypoint localization, its slightly higher P-MPJPE (57.48 mm) suggested limitations in capturing the overall pose configuration. This may stem from an over-emphasis of local features, leading to an inadequate understanding of global pose structure. While the PRM module aimed to balance local and global features, its placement at the final stage of the network may limit its ability to compensate for the local focus of the backbone. The challenge for RSN lies in effectively utilizing enhanced global information from deeper network layers while maintaining sensitivity to local details. Unlike HRNet, which achieves global-local information fusion through parallel sub-networks with cross-connected feature branches, the multi-stage cascade architecture of RSN faces difficulties in effectively transmitting and maintaining global information between stages. Although RSN incorporates mechanisms for global and local multi-level feature fusion within each residual step, these connections may be insufficient for effective inter-stage global information transmission. The network struggles to fuse deep-level global features with shallow-level local features and propagate this fused information through subsequent refinement stages. To address this, one potential solution is to enhance global information transmission across the stages, similar to the cross-stage global information linking of MSPN. This approach could ensure more effective transmission and maintenance of global information throughout the network, potentially improving the ability of RSN to capture the overall pose configuration while retaining its strength in local feature representation.

Simple Baseline performed competitively on P-MPJPE (56.72 mm) metrics. Unlike the complex architecture of RSN, which focus on accurate localization of keypoints but may overlook the overall pose configuration, simpler structure of Simple Baseline potentially achieves a better balance between local accuracy and global consistency in pose estimation tasks. This balance likely contributes to its advantages in P-MPJPE. However, the Simple Baseline model exhibited higher MPJPE (87.55 mm) and MPJVE (85.68 mm/s) compared to more intricate architectures such as MSPN and RSN. Although the

simple architecture may offer an improved balance between local and global features and has advantages in computational efficiency (Xiao et al., 2018), it struggled with the challenges prevalent in complex sports scenarios. Addressing these challenges require sophisticated model architectures capable of more nuanced feature extraction and integration.

HRNet presented the highest MPJPE (87.62 mm) and MPJVE (86.22 mm/frame). Despite its excellent performance in various computer vision tasks (Liu et al., 2022), this model encountered limitations in complex sports scenarios. Its multi-parallel branch structure, designed to maintain high-resolution features and facilitate frequent cross-resolution information exchange, proved crucial for precise keypoint localization (Sun et al., 2019). However, in fast movements, owing to image motion blur, the parallel structure's independent processing of feature information at each branch of the parallel structure can lead to spatial inconsistencies in the features. High-resolution branches may capture the blurred local features, whereas low-resolution branches retain more stable global features that are unaffected by blurring. This disparity can result in spatial misalignment during feature fusion, ultimately reducing the localization accuracy. Conversely, models with serial structures, such as MSPN and RSN, employ progressive down sampling and up sampling process. This stepwise process maintained spatial correspondence of features throughout the network, avoiding the feature fusion misalignment problem in the HRNet architecture. Moreover, utilizing feature skip connections between the same level in down sampling and up sampling can also improve the sophistication of the feature alignment. Furthermore, HRNet consistently maintains high-resolution features throughout its architecture. These high-resolution features were highly susceptible to background interference and partial occlusions, which may also lead to reduced accuracy in the keypoints localization of HRNet, which is highly dependent on these high-resolution features. In addition, during rapid and continuous movements, the high-resolution features can become unstable due to motion blur. HRNet's reliance on these volatile features may also lead to jittery localization results, thus contributing to its poor performance on MPJVE.

Our results from complex sports scenarios, particularly athlete head impacts, demonstrated that the choice of 2D pose estimation method significantly influenced the overall 3D pose estimation performance. The MSPN-based model is suitable for applications requiring high accuracy in keypoint localization within complex sports scenarios

involving rapid movements, partial occlusions, background interference and complex postures. RSN has strength in absolute keypoint localization within complex sports scenarios but has potential limitations in capturing the overall pose configuration. The Simple Baseline model is an efficient choice for applications that focus on capturing the overall structures of human poses. However, these models, along with the HRNet models, exhibited suboptimal performance in terms of absolute keypoint localization accuracy and temporal consistency. Consequently, their applicability to complex sports scenarios remains limited.

In addressing the challenges of complex sports scenarios, different model architectures offer distinct advantages. MSPN and RSN, with their multi-stage cascading and intermediate supervision strategies, along with the effective fusion of local and global features, demonstrated specialized capabilities in handling the spatial and temporal complexity of sports poses. RSN's densely connected structure of RSBs, characterized by multiple branches and convolutional layers, and MSPN's stage-by-stage linking of high-level features from its deep layer network, may contribute to enhanced performance in complex sports scenarios.

Our study proposed improvement strategies for 3D pose estimation models integrating top-down 2D models to address challenges in complex sports scenarios. However, several limitations should be acknowledged. Primarily, our performance evaluation was conducted on a specific dataset, which may limit the generalizability of our results to different sports scenarios or types. Future research should expand the scope to investigate model performance across a wider range of sports activities and examine the impact of diverse training datasets on model performance. Furthermore, this study did not delve into the computational efficiency of the evaluated models. Given the real-time processing requirements common in sports applications, future work should analyse the trade-off between accuracy and computational cost. This analysis could lead to the development of strategies for model compression or optimization techniques, aiming to enhance real-time performance while maintaining model accuracy.

In addition, our study focused exclusively on 3D pose estimation models integrating top-down 2D approaches, which offer high accuracy through specialized single-person pose estimation techniques and robustness to noisy backgrounds by isolating individuals. However, this approach has limitations, particularly in handling occlusions within the bounding box of a target person. In contrast, bottom-

up 2D models, which we did not investigate, offer potential advantages in dealing with partial occlusions. These models rely less on a complete understanding of the entire scene, instead employing a part-to-whole reasoning approach. This methodology allows for gradual construction of the overall pose understanding based on visible local features, potentially yielding more complete representations even when occlusions are present. Thus, it may offer more flexibility in significant occlusion situations due to their ability to piece together available information from visible parts.

Finally, this study focused on a multi-stage framework for 3D human pose estimation, which first estimates 2D poses and then lifts them to 3D. This approach leverages robust 2D pose estimation techniques and performs well in human pose estimations (Liu et al., 2022). However, it has a critical limitation: its heavy reliance on the accuracy of 2D pose estimation. Significant errors in the 2D stage are difficult to correct in the subsequent 3D lifting process, even with robust algorithms. As deep learning and computer vision techniques advance, single-stage methods that predict 3D human poses or body shapes directly from monocular videos are evolving (Mehta et al., 2018; Lin et al., 2023). These methods show promise in overcoming current limitations, potentially reducing model complexity and improving generalization capabilities. Future research directions should include evaluating the integration of bottom-up 2D pose estimation models in the multi-stage framework and comparative analysis of multi-stage and single-stage approaches in complex sports scenarios, especially in athlete head impact events. These studies will provide deeper insights into the strengths and limitations of various model architectures, paving the way for advancements in pose estimation techniques tailored to complex real-world sports scenarios. By exploring these diverse approaches, researchers will be able to work towards more robust, efficient, and accurate pose estimation methods capable of handling the unique challenges presented in dynamic sports environments.

7 CONCLUSION

We aimed to investigate the impact of different top-down 2D pose estimation methods on the performance of a multi-stage 3D pose estimation model in complex sports scenarios, especially in athlete head impact events.

We found that different architectures used for 2D human pose estimation models have different advantages in the 3D human pose estimation task in complex sports scenarios: multi-stage cascading and intermediate supervision (MSPN and RSN), stage-by-stage linking of high-level features in deep layer network (MSPN), fusion of local and global features (MSPN and RSN), and densely connected structure with branches and diverse convolutional layers (RSN). Based on these findings, we concluded that the choice of 2D pose estimation method and their network architectures have a significant effect on the performance of 3D pose estimation in complex sports scenarios, and that different models and architectures are suitable for different application scenarios.

These findings provide strategies for improving 3D pose estimation models and insights and future perspectives for the development of robust and efficient 3D human pose estimation algorithms for complex real-world sports scenarios.

ACKNOWLEDGEMENTS

This work was supported by JST SPRING, Japan Grant Number JPMJSP2106.

REFERENCES

- Liu, W., Bao, Q., Sun, Y., & Mei, T. (2022). Recent advances of monocular 2d and 3d human pose estimation: A deep learning perspective. *ACM Computing Surveys*, 55(4), 1-41.
- Xiao, B., Wu, H., & Wei, Y. (2018). Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 466-481).
- Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.
- Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5693-5703).
- Cai, Y., Wang, Z., Luo, Z., Yin, B., Du, A., Wang, H., ... & Sun, J. (2020). Learning delicate local representations for multi-person pose estimation. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16* (pp. 455-472). Springer International Publishing.
- Li, W., Wang, Z., Yin, B., Peng, Q., Du, Y., Xiao, T., ... & Sun, J. (2019). Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*.
- Bryan, M. A., Rowhani-Rahbar, A., Comstock, R. D., & Rivara, F. (2016). Sports-and recreation-related concussions in US youth. *Pediatrics*, 138(1).
- Giza, C. C., & Hovda, D. A. (2001). The neurometabolic cascade of concussion. *Journal of athletic training*, 36(3), 228.
- Courtney, A., & Courtney, M. (2015). The complexity of biomechanics causing primary blast-induced traumatic brain injury: a review of potential mechanisms. *Frontiers in neurology*, 6, 221.
- McKee, A. C., Stein, T. D., Nowinski, C. J., Stern, R. A., Daneshvar, D. H., Alvarez, V. E., ... & Cantu, R. C. (2013). The spectrum of disease in chronic traumatic encephalopathy. *Brain*, 136(1), 43-64.
- Ji, S., Zhao, W., Ford, J. C., Beckwith, J. G., Bolander, R. P., Greenwald, R. M., ... & McAllister, T. W. (2015). Group-wise evaluation and comparison of white matter fiber strain and maximum principal strain in sports-related concussion. *Journal of neurotrauma*, 32(7), 441-454.
- Camarillo, D. B., Shull, P. B., Mattson, J., Shultz, R., & Garza, D. (2013). An instrumented mouthguard for measuring linear and angular head impact kinematics in American football. *Annals of biomedical engineering*, 41, 1939-1949.
- Madhukar, A., & Ostojic-Starzewski, M. (2019). Finite element methods in human head impact simulations: a review. *Annals of biomedical engineering*, 47(9), 1832-1854.
- Cortes, N., Lincoln, A. E., Myer, G. D., Hepburn, L., Higgins, M., Putukian, M., & Caswell, S. V. (2017). Video analysis verification of head impact events measured by wearable sensors. *The American journal of sports medicine*, 45(10), 2379-2387.
- Camarillo, D. B., Shull, P. B., Mattson, J., Shultz, R., & Garza, D. (2013). An instrumented mouthguard for measuring linear and angular head impact kinematics in American football. *Annals of biomedical engineering*, 41, 1939-1949.
- Wu, L. C., Nangia, V., Bui, K., Hammoor, B., Kurt, M., Hernandez, F., ... & Camarillo, D. B. (2016). In vivo evaluation of wearable head impact sensors. *Annals of biomedical engineering*, 44, 1234-1245.
- King, D., Hume, P. A., Brughelli, M., & Gissane, C. (2015). Instrumented mouthguard acceleration analyses for head impacts in amateur rugby union players over a season of matches. *The American journal of sports medicine*, 43(3), 614-624.
- Li, W., Liu, H., Ding, R., Liu, M., Wang, P., & Yang, W. (2022). Exploiting temporal contexts with strided transformer for 3d human pose estimation. *IEEE Transactions on Multimedia*, 25, 1282-1293.
- Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

- Joher, G., Chaurasia, A., & Qiu, J. (2023). Ultralytics YOLO (Version 8.0.0) [Computer software]. <https://github.com/ultralytics/ultralytics>
- Rogez, G., Weinzaepfel, P., & Schmid, C. (2019). Lcrnet++: Multi-person 2d and 3d pose detection in natural images. *IEEE transactions on pattern analysis and machine intelligence*, 42(5), 1146-1161.
- Moon, G., Chang, J. Y., & Lee, K. M. (2019). Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10133-10142).
- Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7291-7299).
- Newell, A., Yang, K., & Deng, J. (2016). Stacked Hourglass Networks for Human Pose Estimation. *European Conference on Computer Vision*.
- Aharon, N., Orfaig, R., & Bobrovsky, B. Z. (2022). BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*.
- Toshev, A., & Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1653-1660).
- Carreira, J., Agrawal, P., Fragkiadaki, K., & Malik, J. (2016). Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4733-4742).
- Wei, S. E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 4724-4732).
- Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T. S., & Zhang, L. (2020). Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5386-5395).
- Pavlo, D., Feichtenhofer, C., Grangier, D., & Auli, M. (2019). 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7753-7762).
- Pavlakos, G., Zhou, X., Derpanis, K. G., & Daniilidis, K. (2017). Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7025-7034).
- Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2013). Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7), 1325-1339.
- Martinez, J., Hossain, R., Romero, J., & Little, J. J. (2017). A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision* (pp. 2640-2649).
- Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., & Theobalt, C. (2018, September). Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)* (pp. 120-130). IEEE.
- Lin, J., Zeng, A., Wang, H., Zhang, L., & Li, Y. (2023). One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 21159-21168).