PitcherNet: Powering the Moneyball Evolution in Baseball Video Analytics

Jerrin Bright   Bavesh Balaji   Yuhao Chen   David A Clausi   John S Zelek

{jerrin.bright, bbalaji, yuhao.chen1, dclausi, jzelek}@uwaterloo.ca
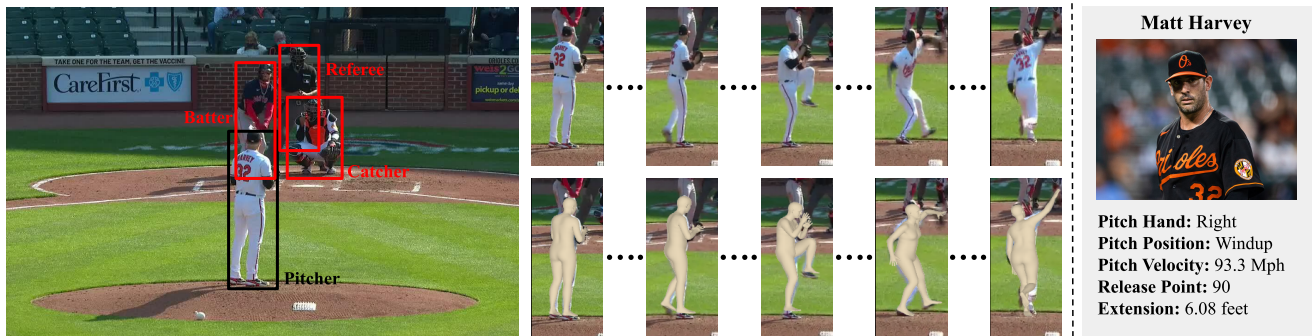University of Waterloo, Waterloo, ON, Canada, N2L 3G1

Figure 1. **3D player reconstruction and kinematic-driven pitch statistics from monocular video.** We introduce *PitcherNet*, a pioneering deep learning system that tackles low-resolution video limitations through efficient 3D human modeling for robust player alignment (left) and reliable pitch statistics analysis from estimated kinematic data (right).

Abstract

In the high-stakes world of baseball, every nuance of a pitcher's mechanics holds the key to maximizing performance and minimizing runs. Traditional analysis methods often rely on pre-recorded offline numerical data, hindering their application in the dynamic environment of live games. Broadcast video analysis, while seemingly ideal, faces significant challenges due to factors like motion blur and low resolution. To address these challenges, we introduce PitcherNet, an end-to-end automated system that analyzes pitcher kinematics directly from live broadcast video, thereby extracting valuable pitch statistics including velocity, release point, pitch position, and release extension. This system leverages three key components: (1) Player tracking and identification by decoupling actions from player kinematics; (2) Distribution and depth-aware 3D human modeling; and (3) Kinematic-driven pitch statistics. Experimental validation demonstrates that PitcherNet achieves robust analysis results with 96.82% accuracy in pitcher tracklet identification, reduced joint position error by 1.8mm and superior analytics compared to baseline methods. By enabling performance-critical kinematic analysis from broadcast video, PitcherNet paves the way for the future of baseball analytics by optimizing pitching strategies, preventing injuries, and unlocking a deeper understanding of pitcher mechanics, forever transforming the game.

1. Introduction

Driven by sabermetrics, pioneered by the Society of American Baseball Research (SABR) [17], baseball analytics has transformed the sport into a data-driven powerhouse, revolutionizing player evaluation, strategic decision-making, and the pursuit of victory [35]. One crucial aspect of this transformation is the analysis of pitch mechanics, where subtle movements such as strides, arm angles, and ball release points significantly affect performance [24, 25]. Analyzing these intricate actions goes beyond traditional pitch type classification (fastball, curveball, etc.), delving into metrics that contribute to strategic deception, such as windup styles, varying velocities, induced ball movement, and ball release point.

Current research on baseball game analysis often rely on numerical databases containing pre-recorded offline data [8, 21, 43, 48]. These methods typically focus on predicting actions or game statistics based on these historical records. While some approaches utilize real-time data, they are often limited to controlled laboratory environments with expensive motion-capture setups [34, 39, 41]. This restricts the generalizability of their findings to the dynamic and complex situations encountered during live games. Live game broadcasts, however, offer a more holistic perspective by capturing the entirety of a pitcher's motion within the game's natural environment. This approach overcomes the limitations of controlled settings. However, analyzing

broadcast data presents its own challenges, such as motion blur and low video resolution, which can significantly hinder accurate pitch analysis and potentially lead to unreliable results.

To bridge the limitations of existing methods and address the challenges of analyzing real-world live broadcasts, we introduce PitcherNet, an end-to-end automated system designed to predict performance-critical pitch statistics from the kinematic data derived from live broadcast videos. PitcherNet transcends existing approaches by meticulously analyzing each stage of the pitcher's movement, from player identification to pose estimation, and finally pitch analysis. Some crucial pitch statistics that PitcherNet estimates include pitch position, pitch velocity, ball release point, and release extension. Human mesh recovery and pitch statistics derived from the PitcherNet system are illustrated in Figure 1. To the best of our knowledge, this is the only system, that extracts pitch statistics extensively driven from the pitcher kinematics from low-quality broadcast videos. The primary contributions of this paper includes the following:

1. We introduce PitcherNet, a novel automated system, which enables accurate prediction of baseball pitch statistics from low-quality broadcast videos.
2. We propose an innovative pitcher identification strategy which aims in classifying players by decoupling actions from player kinematics.
3. Building on D2A-HMR, we improve upon the modeling technique by incorporating transformers with motion blur augmentation and additional regularization heads.
4. We demonstrate state-of-the-art results on a low-quality baseball dataset, MLBPitchDB, validating its applicability in challenging visual environments.

2. Literature Review

Player Tracking and Identification.   Various approaches for player identification exist, primarily relying on either facial features or jersey numbers. In works such as [3, 33], facial recognition is employed to label players based on detected face regions. Conversely, jersey number recognition, as seen in [2, 20, 31, 45, 53], is a prevalent method for player identification. Vats *et al*. [45] recently introduced a comprehensive offline tracking framework for ice hockey, employing 1D convolutions for team and jersey number identification. Sentioscope [4] tracks player interactions using a dual camera setup and model field particles on a calibrated soccer field plane. DeepPlayer [53] proposes a multicamera player identification system integrating jersey number patterns, team classification, and pose-guided partial features. Works such as [2, 13] incorporate end-to-end trainable spatio-temporal networks for identifying jersey numbers in ice hockey and soccer. Additionally, [30, 42, 53] utilize convolutions to extract features and ex-

ploit information from the pose of the players to determine the numbers of the jersey.

Existing player identification approaches rely heavily on distinctive features (such as clothing, jersey number, and facial features). However, these features are often unreliable due to clothing variations, occlusions, and varying camera angles. To address these challenges, we propose a novel approach that decouples player actions from individual tracklets. This approach shifts the focus from specific player features to the action itself, enabling robust and accurate player identification even when traditional features fail.

Player Action Recognition.   Deep learning has emerged as a powerful tool for action recognition, offering promising results. The use of 3D convolutions has demonstrated effectiveness in capturing crucial spatio-temporal information from video data [11, 16, 27, 51]. However, these methods often suffer from a high number of parameters, making them susceptible to overfitting on smaller datasets. To address this limitation, Li *et al*. [27] introduced a spatio-temporal attention network, enabling identification of the key video frames and spatially focus on those frames. Similarly, works including [11, 16, 51, 52] leveraged the pose features from each frame of the sequence and enable effective action recognition without introducing parameter overhead. Yao *et al*. [52] coupled pose and action by formulating pose as an optimization on a set of action-specific manifolds. Cai *et al*. [11] utilize a two-stage architecture that extract pose information and temporal information using optical flow technique before combining them. STAR-Transformer [1] fused video and skeletal data using a transformer architecture with a special cross-attention mechanism. SVFormer [49] introduced a semi-supervised learning approach by incorporating a novel data augmentation technique called Tube TokenMix, specifically designed to improve video understanding.

While recent advancements in player action recognition have been impressive, a gap remains in fully utilizing the rich numerical data embedded within gameplay videos. Integrating this statistical information alongside visual cues holds significant promise for enhancing recognition accuracy and robustness. By incorporating these game statistics, the models can potentially learn more nuanced patterns in player behavior and become less susceptible to outliers.

Baseball Pitch Statistics.   Previous research has mostly focused on estimating the pitch statistics from existing baseball data collection database such as the PITCHf/x system [8, 43, 48]. Works such as [8] leveraged these prior game statistics to classify pitch types using Support Vector Machines (SVM) with linear kernal functions, and [43] utilized Linear Discriminant Analysis (LDA), decision trees, and SVM to find the best apt model to classify pitch types.
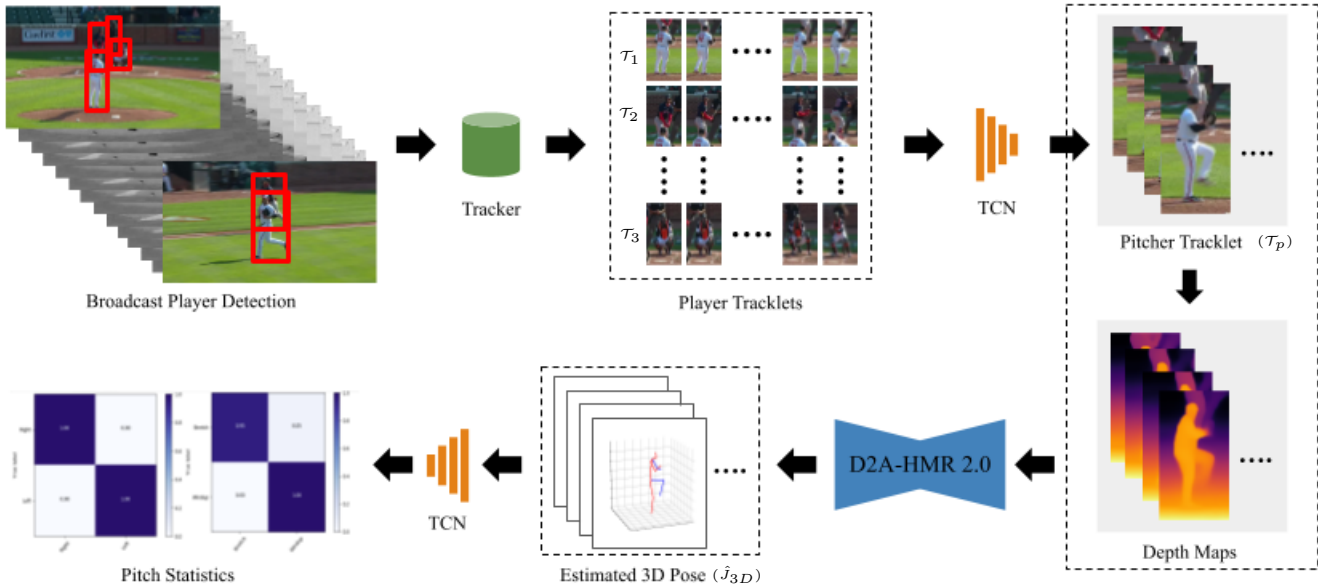
Figure 2. **Overall architecture**. Given a broadcast video, we begin by extracting player tracklets, denoted as $\mathcal{T} \in \{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_n\}$. Each tracklet $\mathcal{T}_k$ consists of a sequence of frames $\mathbf{F}_i$ where $\mathbf{F}_i \in \mathbb{R}^{H \times W \times 3}$ for N frames. These tracklets are then processed through a Temporal Convolutional Network (TCN), which implicitly decouples player actions and identifies the tracklet of the pitcher, called $\mathcal{T}_p$. Subsequently, $\mathcal{T}_p$ undergoes encoding via an encoder ($\mathbf{E}$) to derive pseudo-depth information for each frame. The frames, along with their corresponding pseudo-depth data, are fed into a 3D modeling technique (D2A-HMR 2.0). This framework is responsible for predicting the 3D mesh and 3D joint positions of the pitcher, facilitating detailed analysis of various pitch metrics using the temporal kinematic information processing the 3D joint positions.

Hickey *et al*. [21] aimed to improve the interpretability alongside accuracy of classification models used for pitch prediction. Recently, Manzi *et al*. [34] proposed a descriptive laboratory study in the setting of 3D motion-capture to classify pitch throws by analyzing pitcher kinematics. Similarly, Oyama *et al*. [39] used motion capture systems to validate the pitching motion of the baseball by comparing with the calculated angles. Chen *et al*. [14] proposed a network which can recognize hand pitching style (overhand, three-quarter, etc.) by extracting the human body segment and a descriptor representation using star skeletons.

3. Methodology

The overview of the proposed system, PitcherNet is presented in Figure 2. The system is divided into three components: (1) **Pitcher Tracking and Identification**, involving the initial detection of all players, subsequent tracking of detected players with the assignment of unique labels to each tracklet, and the *decoupling of actions* from the inferred poses of the players in each sequence in the tracklet to facilitate player classification; (2) **3D Human Modeling** utlizes a 3D human model prior [32] to estimate the pose of the player guided by masked modeling, distribution learning, and silhouette masks; and (3) **Pitch Statistics** leveraging TCN and kinematic-driven heuristics to reliably capture various pitch metrics. This section provides a comprehen-

sive exploration of each component, elucidating the underlying design choices that aim to improve the performance of existing techniques in the context of the proposed system.

3.1. Pitcher Tracking and Identification

Accurate tracking and identification of players are fundamental for effective action recognition and analysis in sports scenarios. As highlighted in the literature, the challenges associated with simultaneous tracking and classification based on features are the compromise in reliability of obtaining the desired tracklet. Thus, our objective is to *decouple the action from kinematics obtained from sequences of the tracklets* to acquire the desired tracklet ID for subsequent downstream tasks.

Initially, tracklets are generated using the methodology proposed in SORT [5], which utilize YOLOX [19] detections. Each tracklet is assigned a unique identifier along with the 3D pseudo-pose from MHFormer [28]. MHFormer is chosen due to its lightweight nature which allows efficient processing of the tracklets to estimate player roles. Subsequently, from the pseudo-pose, we decouple player actions by classifying each tracklet into the player's role (pitcher, batter, or others). Given that pitchers are the primary focus of our investigation, we identify sequences within tracklets where pitching actions occur. To accomplish this, we employ a Temporal Convolutional Network (TCN) architecture
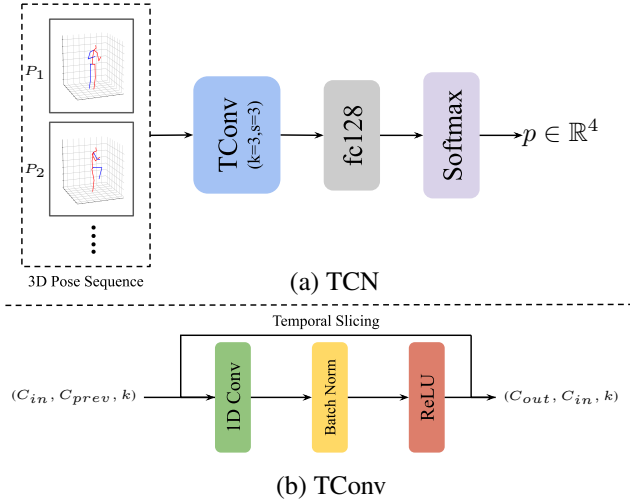
(a) TCN

(b) TConv

Figure 3. **Temporal Convolutional Network.** **(a)** Overview of the proposed TCN for the player identification task, where fc denotes fully connected layers and p refers to the model's output. **(b)** Architecture of the TConv block used in the TCN, where C_{in}, C_{out} and C_{prev} denotes the input, output and previous channels, respectively and k denotes the kernel size.
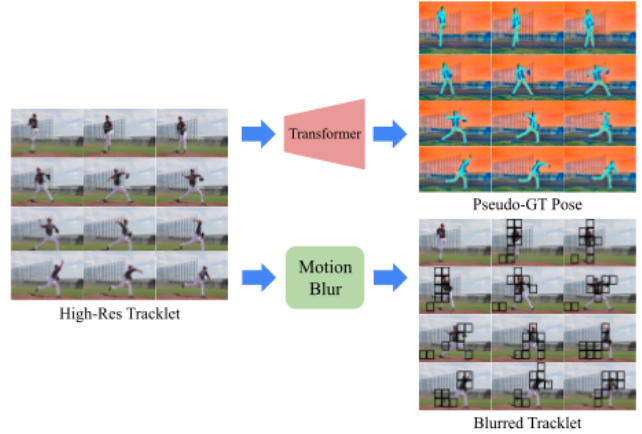

Figure 4. **Data Augmentation Technique.** Pseudo-ground truth pose is collected using a Transformer model for improved generalizability of the pose estimation model.

designed to decouple various actions within each tracklet, specifically isolating the pitching action of interest. The TCN architecture, described in Figure 3, eliminates the dependence on the characteristics of specific players for classification, providing a more robust solution to identify the target player in dynamic sports scenarios.

The TCN architecture utilizes a series of five TConv layers which encompass a dilated 1D convolutional layer with progressively increasing dilation rates, followed by batch normalization and ReLU activation in each layer. The network ingests a 4D tensor representing pose sequences $\mathbb{P} = \{P_i : P \in \mathbb{R}^{K \times C}\}_{i=0}^N$, where each dimension corresponds to batch size (B), temporal sequence length (N), number of joint positions (K), and 3D player coordinates (C). This progressive dilation allows the TCN to capture long-range temporal dependencies crucial for understanding complex motion patterns, while dropout layers and batch normalization enhance the model's generalizability. In addition, skip connections are utilized, allowing the model to directly access information from the original input at a deeper stage in the network. This address the problem of vanishing gradients and improve the flow of information throughout the TCN architecture.

3.2. 3D Human Modeling

Estimating the pose of the pitcher is crucial for effective pitch analysis of the players from live broadcast video. The input to the 3D human modeling technique is the player of interest tracklet from the pitcher tracking and identification component (Section 3.1). To enhance the reliability of pose

estimation in challenging, real-world scenarios, we have advanced the D2A-HMR technique introduced in [10] specifically focusing on our input data conditions.

D2A-HMR focuses on learning the underlying output distribution to minimize the distribution gap. The method takes an input image and pseudo-depth maps, utilizing a transformer encoder that incorporates cross- and self-attention mechanisms along with learnable gate fusions to produce the final output token. Subsequently, a decoder is employed to predict the person's silhouette, guiding the overall structure of the player in the input image. Additionally, a regression head is utilized to obtain the mesh vertices. In this work, we have introduced several design enhancements to augment the D2A-HMR model, and these modifications will be referred to as the D2A-HMR 2.0 modeling technique.

D2A-HMR 2.0 leverages a depth encoder called Depth Anything [50] to extract pseudo-depth, which utilizes the DINOv2 encoder [38] and the DPT decoder [40]. In addition to regression of the mesh vertices as output, we extract the 3D joint coordinates (J_{3D}). Following [15, 23, 26], the mesh vertices are further regressed to find the 3D regressed joint coordinates (J_{3D}^r) using a predefined regression matrix $G \in \mathbb{R}^{K \times M}$. Here K and M correspond to the number of joint positions and the number of vertices. Then, the final 3D pose ($\hat{J}_{3D}$) of the person in the input image is formulated as shown in Equation (1).

$$\hat{J}_{3D} = \omega_1 J_{3D} + (\omega_2 V_{3D})G = \omega_1 J_{3D} + \omega_2 J_{3D}^r \quad (1)$$

where ω_1 and ω_2 are weights for the joint distribution. V_{3D} denotes the vertices of the output mesh.

To further enhance the efficacy of D2A-HMR 2.0, we integrate a substantial volume of unlabeled data sourced

from the Internet to facilitate robust generalization in dynamic real-world scenarios, as illustrated in Figure 4. This augmentation involved the implementation of a transformer network with pretrained weights initialized using MH-Former [28]. Specifically, we select high-resolution practice videos from the Internet and introduce motion blur artifacts. Prior to inducing blur, we use a transformer network to predict its corresponding 3D poses, leveraging its superior performance with high-resolution data.

3.3. Pitch Statistics

The pitch type, a complex interplay of various pitch statistics, ultimately determines the pitch delivered. This work focuses on estimating crucial kinematically derived pitch statistics such as pitch position, release point, release extension, pitch velocity, and handedness from the 3D pose data obtained using the D2A-HMR 2.0 model. While factors such as break and spin rate also influence pitch action [37], this work focuses on these core kinematic statistics mentioned above. By analyzing these statistics, we gain valuable insight into the mechanics of pitch delivery. These pitch statistics combined with kinematic motion data will contribute significantly in the prediction of complex pitch actions. The 3D kinematic information is fed as input to the pitch statistics component to estimate the different pitch statistics, including pitch position, release point, pitch velocity, release extension, and handedness.

3.3.1 Pitch Position

Pitchers utilize two legal pitch position styles: the windup, a full-body motion maximizing power, and the set/stretch, a quicker, more compact motion sacrificing some velocity for faster delivery. Mastering these positions allows pitchers to deceive batters by disrupting their timing and pitch recognition [41]. This work uses a TCN backbone, identical to the player identification network, for pitch position classification using a sigmoid activation function. Each video tracklet is fed into the TCN with a sequence length of 100 frames for classification.

3.3.2 Handedness

Accurate determination of the pitcher's handedness is critical for effective pitch analysis. By isolating the throwing hand within each video frame, we can tailor feature extraction to the pitcher's specific mechanics. This work utilizes a TCN to estimate handedness. While the TCN demonstrates effectiveness, simpler methods based on hand appearance analysis might also be suitable for handedness classification. Regardless of the chosen technique, identifying handedness allows the analysis pipeline to account for the pitcher's mechanics, leading to improved feature extraction and ultimately, more accurate pitch statistics.

3.3.3 Release Point

The release point, defined as the specific location where the pitcher releases the ball from their hand toward the batter, plays a crucial role in determining both the pitch velocity and the release extension [47]. It also plays a crucial factor in deciphering tunneled pitchers.
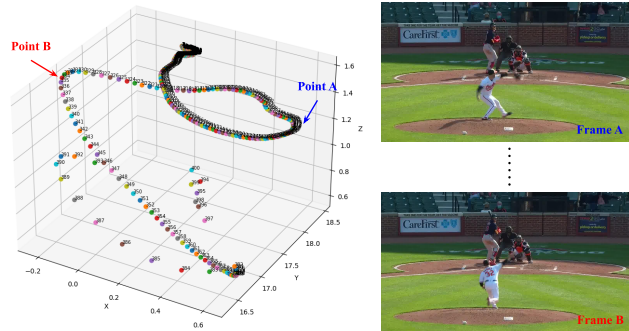

Figure 5. **Trajectory of the right wrist joint in 3D space.** Illustration of two frames which correspond to the points (A and B) marked in the trajectory plot that determines the release point.

As illustrated in Figure 5, we use the wrist kinematics of the pitcher in the x-plane (lateral movement) to identify the release point. Here, we determine the extreme coordinates in the x-plane to establish the maximum and minimum limits of wrist movement during the throwing motion. The limit point, Point A corresponds to the final cocking phase which refers to the point of maximum external rotation of the throwing shoulder from the glove, while the limit point Point B, represents the end of the acceleration phase with the ball release and start of the follow-through phase [36]. We hypothesis that the ball release point will be one amongst n frames windowed on point B with the peak pitch velocity.

3.3.4 Pitch Velocity

Pitch velocity, measured in miles per hour, reflects the ball's speed upon leaving the pitcher's hand [18]. This study proposes a method to estimate pitch velocity by analyzing changes in the throwing hand's wrist position at the release point identified using the estimated 3D pose data. Equation (2) calculates the angular velocity of the wrist at release based on the change in arctangent of consecutive wrist coordinates (w_x, w_y) before (frame $r - 1$) and at (frame r) the release point. This angular velocity is then converted to an approximate pitch velocity (v_p) by multiplying it by the lever arm length (l) between the wrist and the elbow joint.

$$v_p = \omega \times l = \{(atan(w_y^r, w_x^r) - atan(w_y^{r-1}, w_x^{r-1})) \times T\} \times l$$
$$(2)$$

As mentioned above in Section 3.3.3, we estimate the ball release point by finding the maximum velocity in a window around Point B using Equation (2) which will compute the pitch velocity.

3.3.5 Release Extension

Release extension, in baseball, refers to the distance a pitcher creates between the pitching mound and the release point of the ball towards the batter. The release extension helps to differentiate pitch types, as fastballs typically involve greater extension compared to breaking balls. It essentially describes how much closer the pitcher gets to the home plate at the moment of release compared to where they started their throwing motion. The release extension is mathematically indicated as shown in Equation (3).

$$\text{Extension} = \sqrt{(w_x - a_x)^2 + (w_y - a_y)^2 + (w_z - a_z)^2} \tag{3}$$

Here, in Equation (3), a refers to the position of the ankle joint of the pitching leg. The pitching ankle joint is chosen since it remains planted on the mound during the pitch set and the initial part of the delivery.

3.4. Loss Functions

Most pitchers tend to be right-handed. Therefore, there is an imbalance in the class, especially in the handedness data of the pitchers. We use focal loss as a loss function to address the issue of class imbalance [29]. More weight is given to minority classes while training using a gamma tuning parameter (set to 2 initially). The loss function used for the estimation of pitch position and handedness is denoted as shown in Equation (4).

$$L(p_t) = -\alpha_t * (1 - p_t)^\gamma * \log(p_t) \tag{4}$$

where α_t and γ are the balancing parameter and sampling focus parameter, respectively. p_t denotes the predicted probability of the true class. The D2A-HMR human model is trained using the objective mentioned in Equation (5).

$$\mathcal{L}_{model} = \lambda_{RLE}\mathcal{L}_{RLE} + \lambda_{SMPL}\mathcal{L}_{SMPL} + \lambda_{3D}\mathcal{L}_{3D} \\ + \lambda_{2D}\mathcal{L}_{2D} + \lambda_{silh}\mathcal{L}_{silh} \tag{5}$$

where $\mathcal{L}_{RLE}$, $\mathcal{L}_{SMPL}$, $\mathcal{L}_{3D}$, $\mathcal{L}_{2D}$ and $\mathcal{L}_{silh}$ correspond to residual likelihood loss, SMPL vertex loss, regressed 3D loss, reprojected 2D loss and silhouette loss. All λ correspond to the weights assigned to distribute the importance of each objective. We incorporated an additional loss function into our D2A-HMR 2.0 as shown in Equation (6).

$$\hat{\mathcal{L}}_{model} = \mathcal{L}_{model} + \lambda_{3D}^r\mathcal{L}_{3D}^r \tag{6}$$

Here, $\mathcal{L}_{3D}^r$ corresponds to the 3D output joints of the regression head of the D2A-HMR model.

4. Experimentation

Implementation Details. The training process is conducted on three NVIDIA A6000 GPUs with 48GB RAM. Adam optimizer with a batch size of 48 with 500 epochs is used to train the 3D human model. A learning rate of 10^{-4} with betas of 0.9 and 0.99 is used for optimization. The TCN model for handedness estimation and pitch position estimation is trained for 50 and 100 epochs, respectively using one of the three GPUs. The TCN model trained for player identification was trained for 200 epochs using two GPUs with AdamW optimizer with a learning rate of 10^{-2}.

Dataset. We utilize the MLBPitchDB dataset [10] to evaluate our system performance. This dataset is specifically designed for baseball sports analysis and encompasses various data points including player details, 3D pose estimations, actions, and play statistics for all players within the camera's field of view (FoV). Some available ground truth play statistics from the dataset includes pitch velocity, pitch extension, release point of the ball, spin rate, handedness, vertical and horizontal break. Before training, the dataset is pre-processed following the techniques described in [9], which involved player detection, data synchronization, and camera reprojection. To train the 3D human model used for estimating the pose from input images, we utilize the Human 3.6M [22] and 3DPW [46] datasets. In the case of Human3.6M, we trained our D2A-HMR model 2.0 on subjects S1, S5, S6, S7, and S8 and conducted testing on S9 and S11. These configurations were aligned with the common training and evaluation settings within the domain [10, 15].

Metrics. We employ several metrics to assess the performance of different components of our system. For the 3D human pose estimation task, we utilize two primary metrics: the mean per joint position error (mPJPE) and the procrustes-aligned mean per joint position error (PA-mPJPE). These metrics quantify the average distance between the predicted and ground truth 3D joint locations, with PA-mPJPE accounting for global pose variations. For player tracking and identification, we evaluate the system's accuracy performance. Finally, pitch statistics performance is assessed using a standard accuracy metric with different classification margins.

4.1. Pitcher Identification

The impact of the pitcher identification task is compared with two temporal networks (LSTM, transformer with only self-attention blocks) in Table 1. Simple baseline temporal

networks were used for comparison to validate the effectiveness of pose-based role classification. Since these are tasks that use distinct player kinematics, complex networks were not needed.

Table 1. Comparison of our model with baseline temporal networks on MLBPitchDB dataset [10].

	Test Accuracy ↑
LSTM	85.55
Transformer	91.11
Ours	**96.66**

Our approach achieves superior test accuracy compared to both LSTMs and transformers with self-attention blocks, as shown in Table 1. Specifically, we observe an improvement of 11.11% and 5.55% in accuracy relative to LSTMs and transformers, respectively.

4.2. 3D Human Modeling

Depth Encoder. Experimentation with different depth encoders including AdaBin [6], ZoeDepth [7], DINOv1 [12], DINOv2 [38] and Depth Anything [50] is done in Table 2. The D2A-HMR model proposed in [10] utilizes DINOv2 [38] as its depth encoder for human body modeling.

Table 2. Impact on different depth encoders for D2A-HMR evaluated on 3DPW dataset.

	mPJPE ↓	PA-mPJPE ↓
Bhat *et al*. [6]	90.3	55.4
Bhat *et al*. [7]	87.8	53.3
Caron *et al*. [12]	83.1	50.6
Oquab *et al*. [38]	80.5	48.4
Yang *et al*. [50]	**78.7**	**46.9**

Our findings demonstrate that employing Depth Anything [50] as the depth encoder to generate pseudo-depth leads to improvements of 1.8mm and 1.5mm in mPJPE and PA-mPJPE, respectively. This can be attributed to using a significantly larger dataset of unlabeled images allowing the model to learn more comprehensive visual representations. Refer to the Appendix for a qualitative comparison of the various monocular depth estimation techniques on the MLBPitchDB dataset.

Regression Heads. We evaluate the performance of two distinct regressor head architectures within our model. The first design directly regresses the vertex coordinates of the transformer output tokens. On the contrary, the second ap-
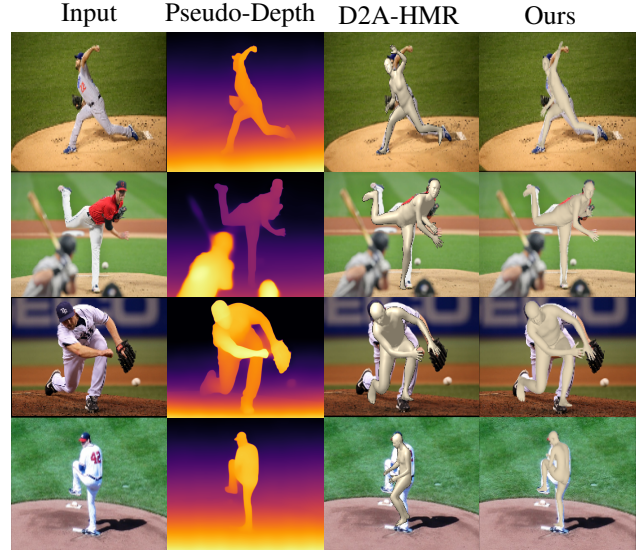

Input   Pseudo-Depth   D2A-HMR   Ours

Figure 6. **Qualitative results.** Qualitative comparison of the pitcher's mesh alignment with the input image using our D2A-HMR 2.0 model.

proach predicts both the vertices and the 3D pose of the players.

Table 3. Ablation study on different regressor heads for D2A-HMR evaluated on 3DPW dataset.

	mPJPE ↓	PA-mPJPE ↓
w/ vertex only	80.5	48.4
w/ vertex+joints	**78.7**	**46.9**

As shown in Table 3, our model demonstrates superior performance when incorporating both the player's vertices and 3D joints during the regression process. This is likely due to the additional information provided by the joints, which helps the model refine the predicted 3D pose and achieve more accurate alignment. Furthermore, the model optimizes the output 3D joints by minimizing the difference between them and the ground truth 3D poses, further contributing to the overall improvement in performance. Figure 6 demonstrates the superior alignment ability of D2A-HMR 2.0 from a given input image when compared with existing state-of-the-art HMR techniques.

Pseudo-GT Data. The impact of using additional pseudo-ground truth pose data on the training process of the D2A-HMR 2.0 model is presented in Table 5. Specifically, pseudo 2D and 3D pose data obtained from HRNet [44] and MHFormer [28], respectively, are used as ground truth to train the HMR model.

Table 4. Performance of our pitch statistics module on different pitch metrics including pitch handedness, pitch position, release point, pitch velocity, and release extension in the test dataset compared against baseline temporal networks.

<div>

(a) Handedness

	Accuracy ↑	F1 Score ↑	Precision ↑
LSTM	85.0	85.7	90.0
Ours (TCN)	**100.0**	**100.0**	**100.0**

(b) Pitch Position

	Accuracy ↑	F1 Score ↑	Precision ↑
LSTM	81.3	82.5	85.0
Ours (TCN)	**97.5**	**97.4**	**95.0**

(c) Release Point

	A_1 ↑	A_2 ↑	A_5 ↑
LSTM	31.3	46.4	63.5
TCN	43.4	51.5	77.6
Ours	**80.8**	**85.8**	**97.9**

(d) Pitch Velocity

	$A_{1\%}$ ↑	$A_{2\%}$ ↑	$A_{5\%}$ ↑
LSTM	5.1	13.1	22.2
TCN	10.1	18.1	48.4
Ours	**43.4**	**68.6**	**94.9**

(e) Release Extension

	$A_{5\%}$ ↑	$A_{8\%}$ ↑	$A_{10\%}$ ↑
LSTM	4.0	7.1	11.1
TCN	14.1	19.1	25.2
Ours	**24.2**	**31.3**	**37.3**

</div>

Table 5. Ablation study on utilizing Pseudo-GT data.

	mPJPE ↓	PA-mPJPE ↓
w/o Pseudo-GT data	79.1	47.4
w/ Pseudo-GT data	**78.7**	**46.9**

Our D2A-HMR 2.0 model demonstrates an improvement in the performance of recovering the 3D mesh by incorporating additional pseudo-ground truth data from the internet, as shown in Table 5.

4.3. Pitch Statistics

Table 4 shows the pitch statistics from the broadcast videos. TCN with five TConv blocks is utilized to predict handedness and pitch position from the kinematic motion sequence of the pitcher. The ball release point is extracted using heuristics from the trajectory of the wrist position of the pitcher and validated using the ground truth release point information from the MLBPitchDB dataset. The pitch velocity and release extension are then computed using mathematical functions that utilize ball release point and kinematic pose information.

Table 4a demonstrates that the TCN model achieves perfect accuracy (100%) in classifying the handedness, without misclassifications for right-handed or left-handed pitchers. The model demonstrates impressive performance in classifying pitch position, correctly identifying 95% of the stretch deliveries and 100% of windup deliveries as shown in Table 4b. The misclassification rate is low, with only 5% of stretch deliveries misclassified as windup, and no misclassification observed for windup deliveries.

The validity of the approach adapted to estimate the release point is examined and compared with alternative methods in Table 4c. The table shows that directly inferring the release point from a temporal network tends to perform poorly. A_x denotes the accuracy with x the number of frames as a margin in the table. Table 4d shows that our

method estimates pitch velocity with superior performance compared to existing temporal networks. $A_{x\%}$ in Table 4d denotes the accuracy with $x\%$ as the margin. Finally, Table 4e presents the superior performance to estimate the release extension. It can be further improved by studying the pitching stride length, which is the distance covered between the spot where one foot hits the ground and the next time the same foot hits the ground again [18]. Our method directly utilizes the release point and 3D pose information to calculate the extension, achieving accurate results compared to the baseline networks.

5. Conclusion

We introduce PitcherNet, an end-to-end deep learning system for kinematic-driven pitch analysis in baseball sports through robust 3D human modeling from broadcast videos. By overcoming challenges such as motion blur in low-resolution feeds, PitcherNet accurately identifies a range of pitch statistics, including pitch position, release point, pitch velocity, pitcher handedness, and release extension. This empowers players, coaches, and fans to gain deeper insights into the technical nuances of pitching and unlock strategic advantages. Additionally, decoupling action from tracklets paves the way for reliable player identification, which holds potential for sports analytics and performance evaluation.

Future work includes extending the framework to analyze further crucial pitch statistics, particularly those involving the pitcher's grip, and delving deeper into player mechanics to optimize posture, reduce injury risk, and further enhance performance.

Acknowledgments

References

[1] Dasom Ahn, Sangwon Kim, Hyunsu Hong, and Byoung Chul Ko. Star-transformer: A spatio-temporal cross attention transformer for human action recognition. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3319–3328, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 2

[2] Bavesh Balaji, Jerrin Bright, Harish Prakash, Yuhao Chen, David A. Clausi, and John Zelek. Jersey number recognition using keyframe identification from low-resolution broadcast videos. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, page 123–130, New York, NY, USA, 2023. Association for Computing Machinery. 2

[3] Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, and Walter Nunziati. Soccer players identification based on visual local features. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, page 258–265, New York, NY, USA, 2007. Association for Computing Machinery. 2

[4] Sermetcan Baysal and Pınar Duygulu. Sentioscope: A soccer player tracking system using model field particles. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(7):1350–1362, 2016. 2

[5] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. 3

[6] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 7

[7] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 7

[8] Joel R Bock. Pitch sequence complexity and long-term pitcher performance. *Sports*, 3(1):40–55, 2015. 1, 2

[9] Jerrin Bright, Yuhao Chen, and John Zelek. Mitigating motion blur for robust 3d baseball player pose modeling for pitch analysis, 2023. 6

[10] Jerrin Bright, Bavesh Balaji, Harish Prakash, Yuhao Chen, David A Clausi, and John Zelek. Distribution and depth-aware transformers for 3d human mesh recovery. *arXiv preprint arXiv:2403.09063*, 2024. 4, 6, 7

[11] Zixi Cai, Helmut Neher, Kanav Vats, David A Clausi, and John S. Zelek. Temporal hockey action recognition via pose and optical flows. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2543–2552, 2018. 2

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 7

[13] Alvin Chan, Martin D. Levine, and Mehrsan Javan. Player

[14] Hua-Tsung Chen, Chien-Li Chou, Wen-Jiin Tsai, Suh-Yin Lee, and Jen-Yu Yu. Extraction and representation of human body for pitching style recognition in broadcast baseball video. In *2011 IEEE International Conference on Multimedia and Expo*, pages 1–4, 2011. 3

[15] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 769–787. Springer, 2020. 4, 6

[16] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. Potion: Pose motion representation for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7024–7033, 2018. 2

[17] Society for American Baseball Research. A guide to sabermetric research. *SABR*, 2021. 1

[18] Dave Fortenbaugh, Glenn Fleisig, and James Andrews. Baseball pitching biomechanics in relation to injury risk and performance. *Sports health*, 1:314–20, 2009. 5, 8

[19] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 3

[20] Sebastian Gerke, Karsten Müller, and Ralf Schäfer. Soccer jersey number recognition using convolutional neural networks. *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 734–741, 2015. 2

[21] Kevin Hickey, Lina Zhou, and Jie Tao. Dissecting moneyball: Improving classification model interpretability in baseball pitch prediction. 2020. 1, 3

[22] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 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, 2014. 6

[23] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 4

[24] Jaret M. Karnuta, Bryan C. Luu, Heather S. Haeberle, Paul M. Saluan, Salvatore J. Frangiamore, Kim L. Stearns, Lutul D. Farrow, Benedict U. Nwachukwu, Nikhil N. Verma, Eric C. Makhni, Mark S. Schickendantz, and Prem N. Ramkumar. Machine learning outperforms regression analysis to predict next-season major league baseball player injuries: Epidemiology and validation of 13,982 player-years from performance and injury profile trends, 2000-2017. *Orthopaedic Journal of Sports Medicine*, 8(11): 2325967120963046, 2020. PMID: 33241060. 1

[25] Matt Kelly. What is sabermetrics? modern analytics impact nearly every part of today's game. *MLB.com*, 2019. 1

[26] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of*

identification in hockey broadcast videos. *Expert Systems with Applications*, 165:113891, 2021. 2

the IEEE/CVF international conference on computer vision, pages 2252–2261, 2019. 4

[27] Dong Li, Ting Yao, Ling-Yu Duan, Tao Mei, and Yong Rui. Unified spatio-temporal attention networks for action recognition in videos. *IEEE Transactions on Multimedia*, 21(2): 416–428, 2019. 2

[28] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. *CoRR*, abs/2111.12707, 2021. 3, 5, 7

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6

[30] Hengyue Liu and Bir Bhanu. Pose-guided r-cnn for jersey number recognition in sports. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2457–2466, 2019. 2

[31] Hongshan Liu, Colin Aderon, Noah Wagon, Huapu Liu, Steven MacCall, and Yu Gan. Deep learning-based automatic player identification and logging in american football videos. *arXiv preprint arXiv:2204.13809*, 2022. 2

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 3

[33] Zahid Mahmood, Tauseef Ali, Shahid Khattak, Laiq Hasan, and Samee U. Khan. Automatic player detection and identification for sports entertainment applications. *Pattern Anal. Appl.*, 18(4):971–982, 2015. 2

[34] Joseph E Manzi, Brittany Dowling, Spencer Krichevsky, Nicholas LS Roberts, Suleiman Y Sudah, Jay Moran, Frank R Chen, Theodore Quan, Kyle W Morse, and Joshua S Dines. Pitch-classifier model for professional pitchers utilizing 3d motion capture and machine learning algorithms. *Journal of Orthopaedics*, 49:140–147, 2024. 1, 3

[35] Joshua Mizels, Brandon Erickson, and Peter Chalmers. Current state of data and analytics research in baseball. *Current reviews in musculoskeletal medicine*, 15(4):283–290, 2022. 1

[36] Tricia A Murray, Timothy D Cook, Sherry L Werner, Theodore F Schlegel, and Richard J Hawkins. The effects of extended play on professional baseball pitchers. *The American journal of sports medicine*, 29(2):137–142, 2001. 5

[37] Daiki Nasu and Makio Kashino. Impact of each release parameter on pitch location in baseball pitching. *Journal of sports sciences*, 39(10):1186–1191, 2021. 5

[38] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4, 7

[39] Sakiko Oyama, Araceli Sosa, Rebekah Campbell, and Alexandra Correa. Reliability and validity of quantitative video analysis of baseball pitching motion. *Journal of applied biomechanics*, 33:64–68, 2017. 1, 3

[40] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021. 4

[41] Donna Moxley Scarborough, Nicholas K Leonard, Lucas W Mayer, Luke S Oh, and Eric M Berkson. The association of baseball pitch delivery and kinematic sequence on stresses at the shoulder and elbow joints. *Journal of Sports Science & Medicine*, 20(1):94, 2021. 1, 5

[42] Arda Senocak, Tae-Hyun Oh, Junsik Kim, and In So Kweon. Part-based player identification using deep convolutional representation and multi-scale pooling. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1813–18137, 2018. 2

[43] Glenn Sidle and Hien Tran. Using multi-class classification methods to predict baseball pitch types. *Journal of Sports Analytics*, 4(1):85–93, 2018. 1, 2

[44] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 7

[45] Kanav Vats, Pascale Walters, Mehrnaz Fani, David A. Clausi, and John S. Zelek. Player tracking and identification in ice hockey. *Expert Systems with Applications*, 213: 119250, 2023. 2

[46] 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)*, pages 601–617, 2018. 6

[47] David Whiteside, Douglas N Martini, Ronald F Zernicke, and Grant C Goulet. Ball speed and release consistency predict pitching success in major league baseball. *The Journal of Strength & Conditioning Research*, 30(7):1787–1795, 2016. 5

[48] Noah Woodward. A decision tree approach to pitch prediction. *The Hardball Times*, 2014. 1, 2

[49] Zhen Xing, Qi Dai, Han Hu, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Svformer: Semi-supervised video transformer for action recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18816–18826, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 2

[50] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891*, 2024. 4, 7

[51] Angela Yao, Juergen Gall, Gabriele Fanelli, and Luc Van Gool. Does human action recognition benefit from pose estimation?". In *Proceedings of the 22nd British machine vision conference-BMVC 2011*. BMV press, 2011. 2

[52] Angela Yao, Juergen Gall, and Luc Van Gool. Coupled action recognition and pose estimation from multiple views. *Int J Comput Vis*, 100:16–37, 2012. 2

[53] Ruiheng Zhang, Lingxiang Wu, Yukun Yang, Wanneng Wu, Yueqiang Chen, and Min Xu. Multi-camera multi-player tracking with deep player identification in sports video. *Pattern Recognition*, 102:107260, 2020. 2