# GolfPose: From Regular Posture to Golf Swing Posture

Ming-Han Lee, Yu-Chen Zhang, Kun-Ru Wu, and Yu-Chee Tseng

Department of Computer Science
National Yang Ming Chiao Tung University
No.1001 University Road, Hsinchu, Taiwan
{mhlee.cs09, yuchen2856.cs10, wufish, yctseng}@nycu.edu.tw

**Abstract.** While there already exist a number of 2D and 3D pose estimation models with high accuracy, in special domains like sports, which usually require even higher accuracy, there are still spaces to be improved. Existing pose models primarily focus on regular daily activities, which, when being applied to precision sports, such as golf swings, still face limitations. In fact, the rare poses and self-occlusions in golf swing videos can easily mislead regular pose models. To overcome these challenges, we develop a small (2D and 3D) GolfSwing dataset that includes both golfer and club poses. We then fine-tune state-of-the-art 2D and 3D posture models, including HRNet, ViTPose, DEKR, and MixSTE, by GolfSwing into a set of models called GolfPose for golfer-club pose estimation with much higher accuracy. Such a simple-yet-effective method may be generalized to other sports with self-occluded properties. Code is available at https://github.com/MingHanLee/GolfPose.

**Keywords:** Human Pose Estimation · Golf · Motion Capture · Precision Sports · Self Occlusion.

## 1 Introduction

Human pose estimation (HPE) has been intensively studied in computer vision and sensor fields. Solutions can be categorized as 2D and 3D ones. Image-based 2D HPE models are proposed in[2,31,8], while 3D postures can be derived by regression [32,12,17] or by 2D-to-3D lifting [28,25]. There are wide ranges of pose applications in sports, including using rugby players' poses to evaluate the risk of concussion during a tackle [27], predicting 3D flight trajectory of badminton [20], incorporating a 3D geometry of the scene to enhance the accuracy of 3D HPE [1], and comparing the pose differences between professional and amateur runners using PoseCoach [19].

In this work, we consider the inference of golf swing videos taken by an off-the-shelf RGB camera. Golf has been increasingly popular in recent years. Golf swings directly impact performance. The studies [44,26] utilized motion capture systems to collect golf swing poses and analyze its relation with injuries. References [14,13] used HPE to identify key frames in golf swing videos and assess

the effectiveness of a swing. How to employ deep learning to coach a beginner's swings based on experts' ones is addressed in [15]. A similar study based on motion capturing is in [16]. The GolfDB dataset [23] consists of 1,400 videos of professional golfers' swings with event frames and bounding boxes labeled. However, the dataset is 2-dimensional and lacks pose keypoint annotations.

Our goal is to estimate both 2D and 3D golfer-with-club postures through a normal RGB camera. To the best of our knowledge, there is no dataset containing all such annotations. We first develop a small *GolfSwing* dataset by a high-quality motion capture system, which features ground truth of 3D golfer-with-club keypoints. There are 17 keypoints for golfer and 5 keypoints for club. We further synchronize these information with normal RGB cameras and project these 3D keypoints to 2D ones as the ground truth. *GolfSwing* enables us to derive a set of more accurate 2D and 3D models, called *GolfPose*, to infer golfer-with-club keypoints through regular videos. In particular, we take a fine-tuning approach. First, a number of 2D state-of-the-art HPE models are fine-tuned, including HRNet [31], ViTPose-H [37], and DEKR [8]. Second, we include club keypoints and fine-tune MixSTE [39], the state-of-the-art 2D-3D lifting model. The results may facilitate various downstream golf applications.

We test these 2D and 3D models fine-tuned from *GolfSwing*. Our experimental results indicate that the original 3D MPJPE of MixSTE can be reduced from 109.4 mm to 35.6 mm and, if we further include club with golfer, the 3D MPJPE can be reduced to a 32.3 mm. For the 2D case, the original mAP of the tested 2D models can be increased from the range of 0.669-0.706 to 0.877-0.936; if we further include club keypoints, the mAP can be increased to 0.918-0.956. This simple-yet-effective approach not only validates the value of *GolfSwing*, but also indicates the feasibility of pretraining a 2D/3D pose model on large datasets like Human3.6M [11] and TotalCapture [33], which primarily focus on regular daily poses, followed by fine-tuning it with a small human-with-object dataset. The results can also be generalized to other precision sports that suffer serious self-occlusion effects, like tennis, badminton, and cricket.

## 2 Related Work

**Motion Capture Systems**. They can be categorized as marker-based, markerless, and inertial sensor-based. Marker-based systems [34] utilize reflective materials to facilitate tracking. Through multiple cameras, the 3D locations of markers are positioned by triangulation. While accurate, such systems are more costly and difficult to set up. Markerless systems [3] do not require markers and track the optical flows of pixels in 2D image spaces for constructing 3D positions. Inertial sensor-based solutions are less costly and provide more degrees of freedom [30]. However, error accumulation is a persistent problem.

Our *GolfSwing* dataset was recorded concurrently by RGB cameras and Vicon cameras (a marker-based system), thus featuring both 2D and 3D ground truth. Markers are attached to both golfer and club. We follow the configurations in [11] in our setup.

**Golf Kinematics.** A lot of studies tried to understand golf kinematics. To study golf swings and injuries, [44] recorded LPGA and PGA golfers' motions and collected statistics including angles and angular velocities of swings. The differences in injury risks and swing techniques among male and female professional golfers on injury regions were studied. The correlation between lumbar and hip joint rotation during a swing and its association with lower back pain was investigated in [26]. To help beginners to correct their poses, HRNet with Simplebaseline3d was employed in [15] to infer 3D poses in GolfDB [23]. Through [16], a learner's poses can be synchronized with coaches' in database, thus providing visualization assistance to learners.

**2D HPE.** 2D HPE can be broadly categorized into two approaches: *top-down* and *bottom-up*. The top-down approach [35,31,37] consists of two stages: object detection and pose estimation. It transforms multi-person pose estimation into single-person estimation. It typically achieves higher accuracy but incurs higher computing cost. The bottom-up approach [2,8,38] first estimates all keypoints, followed by poses construction. This approach is faster, but generally less accurate.

**3D HPE.** Monocular 3D HPE has been widely explored. Solutions can be categorized as one- and two-stage ones. The one-stage approaches [12,17] directly regress 3D skeletons from input without intermediate 2D skeleton representations and are thus more computing-intensive. Two-stage approaches first employ a 2D pose detector to identify skeletons and then elevate 2D skeleton sequences to 3D ones. References [36,41] try to predict 3D skeletons directly from 2D skeletons, and are thus highly sensitive to 2D detection accuracy. Since temporal information of continuous skeletons may reduce depth ambiguity, TCN [28] conducts dilated convolutions on adjacent 2D skeletons to estimate 3D ones. PoseFormer [43] proposes a spatial-temporal transformer encoder to capture skeleton structure and temporal activity. Also based on transformer, MixSTE[39] focuses on the temporal features of individual keypoints and spatial features in each 2D skeleton. Following the recent trend, our GolfPose takes a two-stage approach.

**Human Pose Datasets.** Consisting of 200,000 images and 250,000 person instances, COCO Keypoints [18] defines 17 2D human keypoints and includes annotations for occlusion situations, enabling significant progress under challenging conditions. For 3D datasets, Human3.6M [11] contains large indoor scenarios with 15 daily actions. Also with 17 human keypoints, it includes various data types such as RGB images, human silhouette, bounding box, depth, 3D pose, and 3D laser-scanned human models. MPI-INF-3DHP [24] incorporates both indoor and outdoor scenes with diverse human poses, clothing, and occlusions. TotalCapture [33] provides human keypoints, activity types, and synchronized sensor data. 3DPW [22] is a 3D dataset collected from handheld cameras with sensors attached to human limbs in outdoor environments. SportsPose [10] consists of five types of sports in dynamic scenes.

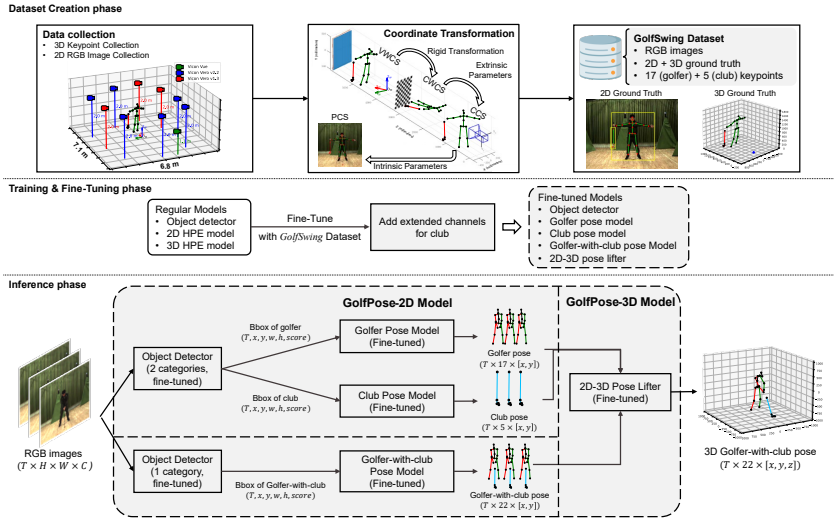


**Fig. 1.** Framework of *GolfPose*. (Blocks marked by gray represent our contributions.)

## 3 Methodology

Our goal is not only to enhance the performance of existing 2D and 3D HPE models but also to include keypoints of club. Fig. 1 shows our research framework. There are three phases. The first phase is to derive *GolfSwing* by Vicon cameras [34] plus regular RGB cameras. In the second phase, we will fine-tune detectors and pose models. The third phase is, from normal RGB videos, to infer golfer-with-club keypoints and, for the case of 3D, to conduct 2D-to-3D lifting.

### 3.1 GolfSwing Dataset

*GolfSwing* is collected concurrently by 9 Vicon infrared cameras and 2 RGB cameras that are time-synchronized. The environment setup and equipment specifications are shown in Fig. 1. The infrared cameras are placed around the golfer, while the RGB cameras are placed in the front and the side of the golfer. The area size is about 6.8m x 7.1m. The golfer stands within the capture region, utilizing a 7-iron club.

Before recording, we calibrate all Vicon cameras with a Vicon wand. A center point on the ground is regarded as the 3D origin. There are 6 volunteer students serving as golfer. Each volunteer is tagged by 28 markers for 3D trajectory tracking. The marker placement is designed similar to Human3.6M, from which we can calculate 17 keypoints as ground truth. In addition, club is tagged by 5 markers for keypoint tracking. The details are depicted in Fig. 2.

Post-processing is required because a marker has to be captured by at least two Vicon cameras in order to reconstruct its 3D location. Due to the speciality of golf sports, markers can be easily occluded during a swing. Missing markers

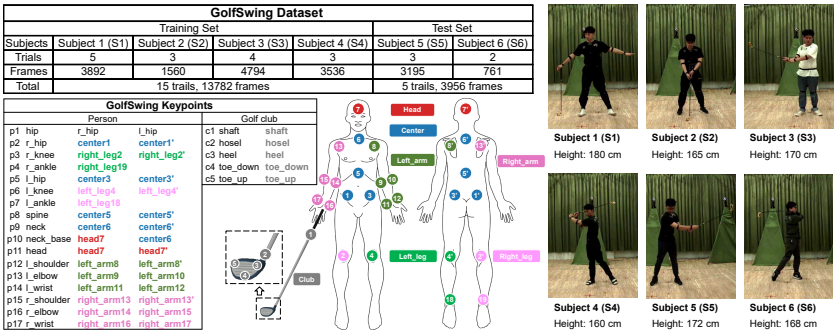


**Fig. 2.** The specifications of *GolfSwing*.

are replaced using Vicon Nexus's algorithms. In the end, we obtain a set of highly accurate 3D golf swing keypoints as ground truth.

The above steps have led to 3D keypoint ground truth. The last step is to perform coordinate transformation to produce 2D keypoint ground truth. This is done by projecting 3D keypoints onto RGB images. We follow Zhang's calibration algorithm [40] and define four coordinate systems (Fig. 1):

1. *Vicon World Coordinate System (VWCS)*: the 3D coordinate system of Vicon cameras, with the calibration wand as the origin.
2. *Checkerboard World Coordinate System (CWCS)*: the 3D coordinate system to relate real world with RGB cameras via an external checkerboard.
3. *Camera Coordinate System (CCS)*: the 3D coordinate system used by RGB cameras.
4. *Pixel Coordinate System (PCS)*: the 2D coordinate system of RGB images, with the top-left corner as the origin.

Then we conduct three coordinate transformations. The first one is VWCS-to-CWCS transformation. We place Vicon markers at the origin, $x$-axis, and $y$-axis of *CWCS* as the transformation basis for *VWCS*. By these markers, we calculate the rotation matrix $R' \in \mathbb{R}^{3\times3}$ and translation matrix $T' \in \mathbb{R}^{3\times1}$, which lead to the Rigid Transformation Matrix $C \in \mathbb{R}^{4\times4}$:

$$C = \begin{bmatrix} R'_{3\times3} & T'_{3\times1} \\ 0_{1\times3} & 0_{1\times1} \end{bmatrix} \tag{1}$$

The second one is CWCS-to-CCS conversion. The Extrinsic Matrix $E \in \mathbb{R}^{4\times4}$ is employed [40]. It is also composed of a rotation matrix $R \in \mathbb{R}^{3\times3}$ and a translation matrix $T \in \mathbb{R}^{3\times1}$, and can be denoted by:

$$E = \begin{bmatrix} R_{3\times3} & T_{3\times1} \\ 0_{1\times3} & 0_{1\times1} \end{bmatrix} \tag{2}$$

The third one is CCS-to-PCS transformation. We utilize the Intrinsic Matrix $K \in \mathbb{R}^{3\times4}$, which consists of the focal length $(f_x, f_y)$ and principal point $(c_x, c_y)$.

It is computed during the calibration algorithm [40]:

$$K = \begin{bmatrix} f_x & 0 & c_x & 0 \\ 0 & f_y & c_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{3}$$

By combining the above transformation matrices, we derive the projection from a 3D point onto the 2D PCS:

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = KEC \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix}, \tag{4}$$

where $\begin{bmatrix} x_w & y_w & z_w & 1 \end{bmatrix}^T$ is a 3D point in *VWCS*, $\begin{bmatrix} u & v & 1 \end{bmatrix}^T$ is its corresponding 2D point in PCS, and $s$ is the scale factor referring to the ratio of the physical measurement unit to the image unit. By projecting all 3D keypoints to PCS, we obtain 2D keypoint ground truth of *GolfSwing*. From the above 2D keypoints, we further calculate the bounding boxes of golfer and club as ground truth.

Overall, *GolfSwing* comprises 6 golfers of heights 160-180 cm, who had taken 2-4 sports classes or joined school sports teams. Each subject swung 7 times, yielding a total of 42 trails. We asked volunteers to swing differently each time. After manual curation, 20 highly accurate trails were collected. We followed "cross-subject" split, with 4 for training and 2 for testing. 17,738 frames were collected, with 13,782 (78%) for training and 3,956 (22%) for testing.

To summarize, there are several challenges during data collection: (i) players' diversity, (ii) recording environments, (iii) fast-moving swings, (iv) missing keypoints in Vicon videos, and (v) opt-in permission required for each volunteer. These are conquered by asking players to swing different each time, cleaning blurry frames (especially for club), and manually making up missing keypoints. Unfortunately, recording in the wild is not feasible currently for Vicon's IR cameras.

### 3.2 Model Fine-Tuning

We take a top-down approach [42] for golfer-with-club keypoint detection. With *GolfSwing*, we fine-tune object detector, 2D pose, and 2D-3D lifting models. In particular, there are two alternatives for object and 2D pose detection, one by detecting golfer and club separately and the other by detecting them jointly.

For object detection, we employ Faster R-CNN [29] and YOLOX [7]. Both models are pretrained on the COCO 2017 dataset, which includes 80 object categories. We fine-tune them by *GolfSwing* (2D) dataset. For the separated method, two categories, namely golfer and club, are detected. For the joint method, only one golfer-with-club category is detected.

For 2D pose detection, we employ HRNet [31] and ViTPose-H [37], which are pretrained on the COCO 2017 dataset for 17 human keypoints. We also include DEKR, which is a bottom-up method, for comparison purpose. Refering to Fig. 1, we then fine-tune them into three models by *GolfSwing* (2D).

- For golfer pose, we modify the configuration file according to the Human3.6M keypoint format. The backbones of the above three models are initialized by pre-trained weights, and we retrain their keypoint heads from scratch, leading to 17 keypoints as output.
- For club pose, we also use the backbones of the above three models and load their pretrained weights. Then we modify their prediction heads for 5 club keypoints and fine-tune them by *GolfSwing* (2D). The club keypoints are shaft, hosel, heel, toe down, and toe-up.
- For golfer-with-club pose, the same fine-tuning is executed except that there are 17+5 keypoints from prediction heads as output.

For 2D-3D lifting, there has been extensive research [28,21,43,39]. We choose to fine-tune the state-of-the-art MixSTE [39]. We follow its design and extend the dimensions of the model to include 5 extra club keypoints. The process is shown in Fig. 3(a). The input to the model is a sequence of $T$ 2D poses $\mathcal{X}_{T,G+C} \in \mathbb{R}^{T\times(G+C)\times 2}$, where $G$ and $C$ are the numbers of golfer and club keypoints, respectively. First, we project each keypoint to $d_m$ dimensions, leading to a higher-dimension feature map $\hat{\mathcal{X}}_{T,G+C} \in \mathbb{R}^{T\times(G+C)\times d_m}$. Then, to preserve positional information, the extended spatial embedding matrix $E_{s\text{-}pos\text{-}ext} \in \mathbb{R}^{(G+C)\times d_m}$ and pre-trained temporal embedding matrix $E_{t\text{-}pos} \in \mathbb{R}^{T\times d_m}$ are applied. (Note that the pre-trained embedding $E_{s\text{-}pos} \in \mathbb{R}^{G\times d_m}$, which is trained on human keypoints only, can not be directly fine-tuned.) Therefore, we randomly initialize $E_{s\text{-}pos\text{-}ext}$ for retaining the positional information of both golfer and club during fine-tuning. Subsequently, $\hat{\mathcal{X}}_{T,G+C}$ will be iternately learned for $l$ iterations between Spatial Transformer Block (STB) and Temporal Transformer Block (TTB). Finally, the dimension $d_m$ is reduced to 3 by the Regression Head, leading to a keypoint sequence $\mathcal{Y}_{T,G+C} \in \mathbb{R}^{T\times(G+C)\times 3}$.

The modified STB and TTB transformer blocks are shown in Fig. 3(b). We follow the transformer encoders designed in [6,43,39]. STB is to learn the spatial relationships among keypoints in each frame. Frames are sent one-by-one to STB. With the pre-trained weights of MixSTE, the *multi-head self-attention* of STB already effectively preserved the spatial relation of keypoints for regular human activities. During fine-tuning, for each frame at time $t$, its (dimension-incremented) keypoints, denoted by $i_{t,n} \in \mathbb{R}^{d_m}, n = 1...(G+C)$, are regarded as a sequence of tokens by STB to enhance their spatial relation-capturing capability, such as inter-golfer, inter-club, and golf-club keypoints' relationships. On the other hand, the trajectories of all keypoints along the temporal dimension are also sent one-by-one to TTB. For each trajectory $n$, $n = 1..G+C$, its (dimension-incremented) keypoints, denoted by $i_{n,t} \in \mathbb{R}^{d_m}, t = 1...T$, are regarded as a sequence of tokens by TTB to enhance their temporal relation-capturing capability. Overall, these two blocks alternately strengthen the correlations of keypoints in spatial and temporal dimensions, respectively.

This model is fine-tuned end-to-end in a supervised manner. We adopt the same loss functions: Weight Mean Per Joint Position Error (W-MPJPE) $L_w$ and Mean Per Joint Velocity Error (MPJVE) $L_v$ [28]. Additionally, we adopt Temporal Consistency Loss (TCLoss) $L_c$ to improve motion smoothness [9].

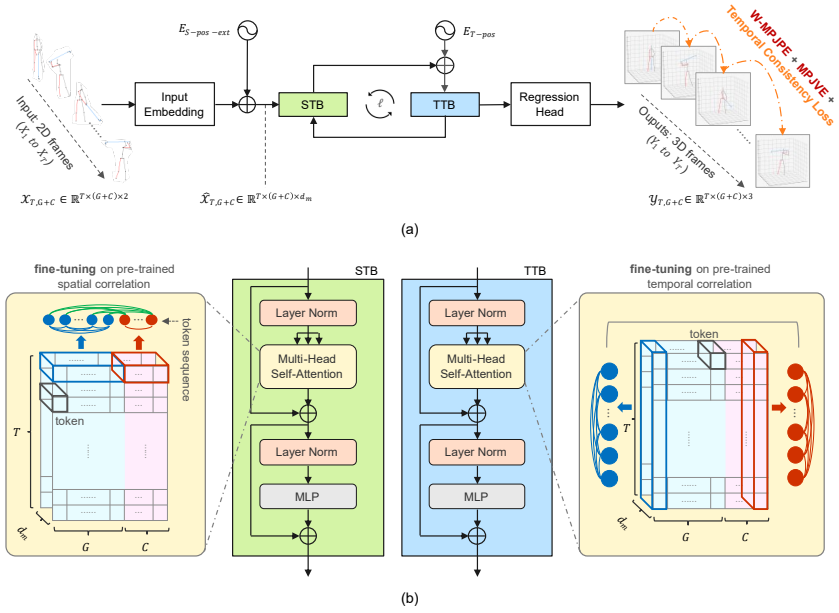


**Fig. 3.** (a) Extension of MixSTE for 2D-3D lifting and (b) extension of STB and TTB transformer blocks to include club keypoints.

### 3.3 GolfPose Inference Model

*GolfPose* is built upon the above fine-tuned models. As shown in Fig. 1, it accepts a RGB frame sequence of length $(T, H, W, C)$ as input. If one chooses to process golfer and club separately, we need to identify from each frame a golfer bounding box $G_b = (p_x, p_y, p_w, p_h, score)$ and a club bounding box $C_b = (c_x, c_y, c_w, c_h, score)$. Then, $P_b$ and $C_b$ are passed to the golfer and the club pose models, respectively. Then golfer and club keypoints of all $T$ frames are stacked into tensors of $(T, 17, 2)$ and $(T, 5, 2)$, respectively, which are then concatenated into a $(T, 22, 2)$ tensor. If one chooses to process golfer and club jointly, the process is similar, except that there is only one bounding box per frame and we directly derive a $(T, 22, 2)$ tensor. In either case, the concatenated tensor is fed into the 2D-3D lifter. With joint golfer-with-club information, we shall show that the lifter can better leverage the spatial-temporal correlations of keypoints and thus achieve much higher accuracy.

## 4 Performance Evaluation

### 4.1 Implementation Details

As mentioned earlier, our 2D/3D models are pre-trained on the COCO 2017 dataset and the Human3.6M dataset, respectively, and then fine-tuned on *GolfSwing* 2D/3D. For *GolfSwing*, the training set (S1-S4) consists of 13,782 images,

**Table 1.** Comparisons of 3D pose estimation models on our *GolfSwing* 3D dataset.

| Strategy | w/o Fine-tuned | with Fine-tuned | |
|---|---|---|---|
| | Golfer | Golfer | Club |
| VideoPose3D [28] (N=17, T=243) | 134.8 | 52.0 | - |
| Attention3D [21] (N=17, T=243) | 149.7 | 46.8 | - |
| PoseFormer [43] (N=17, T=81) | **107.8** | 40.3 | - |
| MixSTE [39] (N=17, T=243) | 109.4 | **35.6** | - |
| GolfPose-3D(GC) (N=22, T=243) | - | **32.3** | 62.8 |

2D/3D keypoints, and 27,564 bounding box annotations, while the test set (S5, S6) consists of 3,956 images, 2D/3D keypoints, and 7,912 bounding box annotations. For object detectors, the evaluation metric is $mAP@IoU$. During fine-tuning, we set a batch size of 8 and train the models for 30 epochs. The optimizer is SGD, and the learning rate is set to 2.5e-3. The computing environments are: CPU i7-12700K, GPU GeForce RTX 3090*2, CUDA 11.6, PyTorch 1.12.1, and mmdetection [4] version 3.1.0.

For 2D pose models, the evaluation metric is $mAP@OKS$. The computing environments are the same but with additional mmpose [5] version 1.3.0. During fine-tuning, we set the batch size to 16 and train the models for 20 epochs. We use Adam optimizer, with a learning rate of 1e-4. For the golfer's 17 keypoints, we assign different weights [1.0, 1.0, 1.2, 1.5, 1.0, 1.2, 1.5, 1.0, 1.0, 1.0, 1.0, 1.0, 1.2, 1.5, 1.0, 1.2, 1.5] to them when calculating MSE loss. For the club's 5 keypoints, we assign weights [1.6, 1.9, 2.0, 2.0, 2.0] to them. The golfer-with-club's keypoionts are given weights similarly.

To fine-tune MixSTE, in addition to extending to 22 keypoints, we perform data augmentation on *GolfSwing* to enhance robustness. We rotate each 3D pose by 90 degrees and project it onto the 2 RGB cameras. So the dataset quadrupled, effectively rendering additional perspectives of 2D poses. We divide keypoints into five groups (head, torso, upper limbs, lower limbs, and club) and define the weight vector $W = [1.5, 1, 2.5, 4, 4]$. The frame length $T = 243$ and the Adam optimizer is employed with a learning rate of 4.0e-5 and a decay of 0.98 per epoch. The batch size is 512 and the model is fine-tuned for 60 epochs. The computing environments are: CPU i7-12700K, GPU GeForce RTX 3090*2, CUDA 11.6, and PyTorch 1.10.1.

### 4.2 Performance Comparison

**Quantitative Results.** We compare *GolfPose* against four 3D models VideoPose3D [28], Attention3D [21], PoseFormer [43], and MixSTE [39] on the *GolfSwing* dataset. Among them, MixSTE is the current state-of-the-art in Human3.6M. We use the 2D pose ground truth as input to compare the predicted 3D poses by the MPJPE metric. We use the default hyper-parameters of these four models during fine-tuning. As Table 1 shows, these four models all improve

**Table 2.** Comparisons of 2D pose models on our *GolfSwing* 2D dataset. (G, C, and GC mean fine-tuning for golfer only, for club only, and for both, respectively. "Metric" means the range of keypoints in calculating AP and AR.)

| Model | Source model | Metric | AP | $AP^{50}$ | $AP^{75}$ | AR | $AR^{50}$ | $AR^{75}$ |
|---|---|---|---|---|---|---|---|---|
| GolfPose-2D(G) | HRNet | $AP_{golfer}$ $AR_{golfer}$ | 0.884 | **1.000** | **1.000** | 0.887 | **1.000** | **1.000** |
| GolfPose-2D(GC) | | | 0.899 | **1.000** | **1.000** | 0.916 | **1.000** | **1.000** |
| GolfPose-2D(G) | ViTPose-H | | 0.887 | **1.000** | **1.000** | 0.898 | **1.000** | **1.000** |
| GolfPose-2D(GC) | | | 0.901 | **1.000** | **1.000** | 0.915 | **1.000** | **1.000** |
| GolfPose-2D(G) | DEKR | | 0.869 | **1.000** | 0.898 | 0.888 | **1.000** | 0.904 |
| GolfPose-2D(GC) | | | **0.917** | **1.000** | 0.968 | **0.927** | **1.000** | 0.973 |
| GolfPose-2D(C) | HRNet | $AP_{club}$ $AR_{club}$ | 0.857 | 0.990 | 0.947 | 0.882 | 0.997 | 0.957 |
| GolfPose-2D(GC) | | | 0.949 | **1.000** | 0.990 | 0.955 | **1.000** | 0.999 |
| GolfPose-2D(C) | ViTPose-H | | 0.870 | 0.990 | 0.948 | 0.887 | 0.996 | 0.953 |
| GolfPose-2D(GC) | | | 0.942 | **1.000** | 0.990 | 0.956 | **1.000** | 0.998 |
| GolfPose-2D(C) | DEKR | | 0.858 | 0.990 | 0.946 | 0.888 | 0.999 | 0.951 |
| GolfPose-2D(GC) | | | **0.977** | **1.000** | **1.000** | **0.982** | **1.000** | **1.000** |
| GolfPose-2D(GC) | HRNet | $AP_{golfer-club}$ $AR_{golfer-club}$ | 0.915 | **1.000** | **1.000** | 0.930 | **1.000** | **1.000** |
| | ViTPose-H | | 0.925 | **1.000** | **1.000** | 0.930 | **1.000** | **1.000** |
| | DEKR | | **0.942** | **1.000** | **1.000** | **0.945** | **1.000** | **1.000** |
| HRNet [31] | - | $AP_{limb}$ $AR_{limb}$ | 0.701 | **1.000** | 0.948 | 0.731 | **1.000** | 0.954 |
| GolfPose-2D(G) | HRNet | | 0.918 | **1.000** | **1.000** | 0.939 | **1.000** | **1.000** |
| GolfPose-2D(GC) | HRNet | | **0.956** | **1.000** | **1.000** | **0.962** | **1.000** | **1.000** |
| ViTPose-H [37] | - | | 0.706 | **1.000** | **1.000** | 0.730 | **1.000** | **1.000** |
| GolfPose-2D(G) | ViTPose-H | | 0.936 | **1.000** | **1.000** | 0.947 | **1.000** | **1.000** |
| GolfPose-2D(GC) | ViTPose-H | | 0.941 | **1.000** | **1.000** | 0.948 | **1.000** | **1.000** |
| DEKR [8] | - | | 0.669 | **1.000** | 0.979 | 0.689 | **1.000** | 0.988 |
| GolfPose-2D(G) | DEKR | | 0.877 | **1.000** | 0.868 | 0.887 | **1.000** | 0.871 |
| GolfPose-2D(GC) | DEKR | | 0.918 | **1.000** | 0.927 | 0.924 | **1.000** | 0.935 |

significantly after fine-tuning, implying the contribution of *GolfSwing*. After fine-tuning, MixSTE performs the best. Encompassing club information, *GolfPose* generates $N = 22$ keypoints and outperforms the other methods, which only yield $N = 17$ golfer keypoints. This indicates that including object is helpful for pose estimation. After fine-tuning, *GolfPose* achieves the lowest MPJPE of 32.3 mm for golfer keypoints. In fact, the club's MPJPE=62.8 mm because its fast-moving nature causes blurry effects. Even under such a condition, it still proves the importance of including club for golfer pose estimation.

Next, we consider the 2D pose estimation results, including golfer-only, club-only, and golfer-with-club cases. Table 2 presents two types of results: HRNet and ViTPose-H represent the top-down approach, and DEKR represents the bottom-up approach. If we fine-tune for golfer only (G) or for club only (C) by *GolfSwing* 2D, ViTPose-H performs the best with mAP=0.887 and 0.870, respectively (underlined). If we fine-tune for both golfer and club (GC), all models are further improved after fine-tuning. DEKR achieves the highest mAP of 0.917 in golfer's keypoints, of 0.977 in club's keypoints, and of 0.942 in all keypoints (boldface). These results indicate that including club benefits golfer keypoint detection, and reversely including golfer benefits club keypoint detection.

Finally, we compare the pre-trained and the fine-tuned models. Since COCO and Human3.6M define skeleton differently, we have to take the 12 common key-

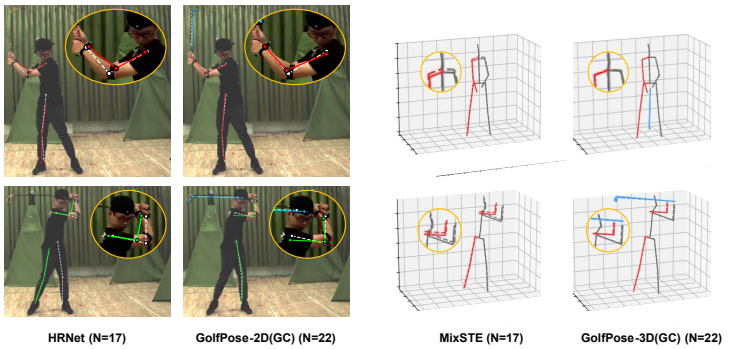


**Fig. 4.** Qualitative comparisons on subject S5. (GT=dashed line; prediction=solid line; red=right hand; green=left hand)

points between them (which include shoulders, elbows, wrists, hips, knees, and ankles). The results are shown in the last section of Table 2. For each source model (HRNet, ViTPose-H, and DEKR), our fine-tuned *GolfPose* does improve mAP significantly. Overall, using golfer-with-club data to fine-tune HRNet performs the best, achieving mAP= 0.956. This implies the value of *GolfSwing* that makes 2D pose estimation more stable and accurate, which can further contribute to the subsequent 2D-3D lifter.

**Qualitative Results.** Fig. 4 presents some qualitative results. The visualization is from S5 of *GolfSwing*. The results show the improvement from the pre-trained model to the fine-tuned model. Notably, even when hands are partially occluded, *GolfPose*-2D and -3D can still detect keypoints quite accurately.

**Inference speed.** Regarding inference speed, *GolfPose* mainly involves a fine-tuned 2D golfer pose model, a fine-tuned 2D club pose model, and a fine-tuned 2D-3D lifter. The first two models, when running on an i5-12500 CPU and GeForce GTX 1080 Ti GPU, achieve 27.25 and 27.3 FPS, respectively. The third model, when running on an i5-13400 CPU with a GeForce RTX 3060 GPU, reaches 6.67 FPS.

### 4.3 Ablation Study

**Object Detection.** Table 3 compares the case of detecting golfer and club separately and the case of detecting them jointly. We test two object detectors: Faster R-CNN and YOLOX-s. There is clear advantage of detecting them jointly. Contrary to intuition, when each individual object's detection is low, jointly detecting them helps improve detection rate. With joint detection, YOLOX-s outperforms Faster R-CNN. Additionally, YOLOX-s boasts a higher inference speed of 88.84 FPS compared to Faster R-CNN's 14.19 FPS. Therefore, we adopt YOLOX-s as our object detector.

**Table 3.** Ablation study on separate and joint object detection.

| Model | Datasets | Class | AP | $AP^{50}$ | $AP^{75}$ |
|---|---|---|---|---|---|
| Faster R-CNN (ResNet50-FPN) | Coco + *GolfSwing* | Golfer | 0.940 | 1.000 | 1.000 |
| | | Club | 0.896 | 1.000 | 0.989 |
| | | G-w-C | 0.970 | 1.000 | 0.990 |
| YOLOX-s (CSPDarknet) | Coco + *GolfSwing* | Golfer | 0.920 | 1.000 | 1.000 |
| | | Club | 0.911 | 1.000 | 0.998 |
| | | G-w-C | 0.984 | 1.000 | 1.000 |

**Table 4.** Ablation study on the number of club keypoints.

| MPJPE (mm) | Number of keypoints | | | | | |
|---|---|---|---|---|---|---|
| | 17+0 | 17+1 | 17+2 | 17+3 | 17+4 | 17+5 |
| Golfer | 35.6 | 29.5 | 30.5 | 30.8 | 32.3 | 32.3 |
| Club | - | 50.9 | 59.6 | 62.9 | 61.4 | 62.8 |
| Overall | 35.6 | 30.7 | 33.6 | 35.6 | 37.9 | 39.2 |

**Table 5.** Ablation study on fine-tuning from Human3.6M.

| MPJPE (mm) | Train from scratch | Fine-tuning |
|---|---|---|
| Golfer | 48.5 | 32.3 |
| Club | 112.9 | 62.8 |
| Overall | 63.2 | 39.2 |

**Number of club keypoints.** In Table 4, we further consider the effect of the number of club keypoints. We denote the method by $17+i$, where $i = 0..5$ represents the number of club keypoints (when $0 < i < 5$, we choose keypoints c1 to c$i$ in Fig. 2). When we start to add club keypoint, we observe significant improvement on both golfer and golfer-with-club detection accuracy (from error=35.6 mm to 29-32 mm for golfer). However, adding more keypoints results in slight increases of error. We suspect the reason to be the relative slower movement of the grip part as opposed to the much faster movement of the head part of club. As mentioned earlier, it is more difficult to detect fast-moving keypoint. Therefore, when the value of $i$ increases, these keypoints (of relative lower accuracy) also confuse our model.

**Effect of fine-tuning.** We consider the same structure of *GolfPose* that is trained from scratch on *GolfSwing* for 80 epochs (i.e., without using the pretrained weights from MixSTE). From Table 5, it validates the benefit of the pretrained weights from MixSTE (which reduces error from 48.5 mm to 32.3 mm for golfer). That is, a large amount of information is carried over from the pre-trained weights obtained from Human3.6M. Even for club, the error is reduced from 112.9 mm to 62.8 mm.

## 5 Conclusions

This work contributes in deriving the *GolfSwing* dataset, which includes keypoint ground truth of 2D and 3D golf swing actions. It also contributes in deriving the *GolfPose* framework, which can be fine-tuned from existing object detection

and pose estimation models, for inferring golfer-with-club keypoints simultaneously. The results imply that including auxiliary objects, such as club, with even very few keypoints of a small dataset can improve human pose estimation significantly. Nonetheless, detecting club poses in complex scenes is a challenge. Future improvement on club pose estimation may further improve overall performance. This approach can be extended to other sports, such as baseball, cricket, badminton, and tennis, where players have an object at hand.

## References


1. Baumgartner, T., Klatt, S.: Monocular 3d human pose estimation for sports broadcasts using partial sports field registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5108–5117 (2023)
2. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7291–7299 (2017)
3. Captury: Captury motion systems. https://captury.com/ (2013), accessed: 2023-06-19
4. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
5. Contributors, M.: Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose (2020)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929 (2020)
7. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
8. Geng, Z., Sun, K., Xiao, B., Zhang, Z., Wang, J.: Bottom-up Human Pose Estimation via Disentangled Keypoint Regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14676–14686 (2021)
9. Hossain, M.R.I., Little, J.J.: Exploiting Temporal Information for 3D Human Pose Estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 68–84 (2018)
10. Ingwersen, C.K., Mikkelstrup, C., Jensen, J.N., Hannemose, M.R., Dahl, A.B.: SportsPose: A Dynamic 3D Sports Pose Dataset. In: Proceedings of the IEEE/CVF International Workshop on Computer Vision in Sports (2023)
11. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: 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 (2013)
12. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end Recovery of Human Shape and Pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7122–7131 (2018)

13. Kim, T.T., Zohdy, M.A., Barker, M.P.: Applying Pose Estimation to Predict Amateur Golf Swing Performance using Edge Processing. IEEE Access **8**, 143769–143776 (2020)
14. Lee, K.J., Ryou, O., Kang, J.: Quantitative Golf Swing Analysis based on Kinematic Mining Approach. Korean Journal of Sport Biomechanics **31**(2), 87–94 (2021)
15. Liao, C.C., Hwang, D.H., Koike, H.: AI Golf: Golf Swing Analysis Tool for Self-Training. IEEE Access **10**, 106286–106295 (2022)
16. Liao, C.C., Hwang, D.H., Wu, E., Koike, H.: AI Coach: A Motor Skill Training System using Motion Discrepancy Detection. In: Proceedings of the Augmented Humans International Conference. pp. 179–189 (2023)
17. Lin, K., Wang, L., Liu, Z.: End-to-end Human Pose and Mesh Reconstruction with Transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1954–1963 (2021)
18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 740–755. Springer International Publishing (2014)
19. Liu, J., Saquib, N., Chen, Z., Kazi, R.H., Wei, L.Y., Fu, H., Tai, C.L.: PoseCoach: A Customizable Analysis and Visualization System for Video-based Running Coaching. In: IEEE Transactions on Visualization and Computer Graphics. pp. 1–14 (2022)
20. Liu, P., Wang, J.H.: MonoTrack: Shuttle Trajectory Reconstruction From Monocular Badminton Video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3513–3522 (2022)
21. Liu, R., Shen, J., Wang, H., Chen, C., Cheung, S.c., Asari, V.: Attention Mechanism Exploits Temporal Contexts: Real-time 3D Human Pose Reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5064–5073 (2020)
22. Timo von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 601–617 (2018)
23. McNally, W., Vats, K., Pinto, T., Dulhanty, C., McPhee, J., Wong, A.: Golfdb: A Video Database for Golf Swing Sequencing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR). pp. 0–0 (2019)
24. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision. In: Proceedings of the International Conference on 3D Vision (3DV) (2017)
25. Mohamed, A., Chen, H., Wang, Z., Claudel, C.: Skeleton-graph: Long-term 3D Motion Prediction from 2D Observations using Deep Spatio-temporal Graph CNNs. arXiv preprint arXiv:2109.10257 (2021)
26. Mun, F., Suh, S.W., Park, H.J., Choi, A.: Kinematic Relationship Between Rotation of Lumbar Spine and Hip Joints during Golf Swing in Professional Golfers. Biomedical Engineering Online **14**, 1–10 (2015)
27. Nonaka, N., Fujihira, R., Nishio, M., Murakami, H., Tajima, T., Yamada, M., Maeda, A., Seita, J.: End-to-End High-Risk Tackle Detection System for Rugby. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3550–3559 (2022)

28. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 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 (CVPR). pp. 7753–7762 (2019)
29. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (Jun 2017)
30. Roetenberg, D., Luinge, H., Slycke, P., et al.: Xsens MVN: Full 6DOF Human Motion Tracking using Miniature Inertial Sensors. Xsens Motion Technologies BV, Tech. Rep **1**, 1–7 (2009)
31. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep High-resolution Representation Learning for Human Pose Estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5693–5703 (2019)
32. Tekin, B., Rozantsev, A., Lepetit, V., Fua, P.: Direct Prediction of 3D Body Poses from Motion Compensated Sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 991–1000 (2016)
33. Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J.: Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors. In: Proceedings of British Machine Vision Conference. pp. 1–13 (2017)
34. Vicon: Motion Capture. https://www.vicon.com/ (1984), accessed: 2023-08-07
35. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019)
36. Xiao, B., Wu, H., Wei, Y.: Simple Baselines for Human Pose Estimation and Tracking. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 466–481 (2018)
37. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation. In: Advances in Neural Information Processing Systems (2022)
38. Yu-Hui, C., Ard, O., Francois, B., Andrew, B., Vijay, S.: MoveNet. https://www.tensorflow.org/hub/tutorials/movenet (2021)
39. Zhang, J., Tu, Z., Yang, J., Chen, Y., Yuan, J.: MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13232–13242 (2022)
40. Zhang, Z.: A Flexible New Technique for Camera Calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(11), 1330–1334 (2000)
41. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic Graph Convolutional Networks for 3D Human Pose Regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3425–3435 (2019)
42. Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., Shah, M.: Deep Learning-based Human Pose Estimation: A Survey. arXiv preprint arXiv:2012.13392 (2020)
43. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3D Human Pose Estimation With Spatial and Temporal Transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11656–11665 (2021)
44. Zheng, N., Barrentine, S., Fleisig, G., Andrews, J.: Swing Kinematics for Male and Female Pro Golfers. International Journal of Sports Medicine **29**(12), 965–970 (2008)