International Journal of Computer Vision manuscript No.
(will be inserted by the editor)

SportsCap: Monocular 3D Human Motion Capture and Fine-grained Understanding in Challenging Sports Videos

Xin Chen[1,2,3] · **Anqi Pang**[1,2,3] · **Wei Yang**[4] · **Yuexin Ma**[1] · **Lan Xu**[1] · **Jingyi Yu**[1]

Received: date / Accepted: date

Abstract Markerless motion capture and understanding of professional non-daily human movements is an important yet unsolved task, which suffers from complex motion patterns and severe self-occlusion, especially for the monocular setting. In this paper, we propose SportsCap – the first approach for simultaneously capturing 3D human motions and understanding fine-grained actions from monocular challenging sports video input. Our approach utilizes the semantic and temporally structured sub-motion prior in the embedding space for motion capture and understanding in a data-driven multi-task manner. To enable robust capture under complex motion patterns, we propose an effective motion embedding module to recover both the implicit motion embedding and explicit 3D motion details via a corresponding mapping function as well as a sub-motion classifier. Based on such hybrid motion information, we introduce a multi-stream spatial-temporal Graph Convolutional Network(ST-GCN) to predict the fine-grained semantic action attributes, and adopt a semantic attribute mapping block to assemble various correlated action attributes into a high-level action label for the overall detailed understanding of the whole sequence, so as to enable various applications like action assessment or motion scoring. Comprehensive experiments on both public and our proposed datasets show that with a challenging monocular sports video input, our novel approach not only significantly improves the accuracy of 3D human motion capture, but also recovers accurate fine-grained semantic action attribute.

Keywords Human modeling · 3D motion capture · Motion understanding

Xin Chen
E-mail: chenxin2@shanghaitech.edu.cn

Anqi Pang
E-mail: pangaq@shanghaitech.edu.cn

Wei Yang
E-mail: wyangcs@udel.edu

Yuexin Ma
E-mail: mayuexin@shanghaitech.edu.cn

Lan Xu
E-mail: xulan1@shanghaitech.edu.cn

Jingyi Yu
E-mail: yujingyi@shanghaitech.edu.cn
[1] Shanghai Engineering Research Center of Intelligent Vision and Imaging School of Information Science and Technology ShanghaiTech University
[2] Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences
[3] University of Chinese Academy of Sciences
[4] Dgene Co.,Ltd.

1 Introduction

The past ten years have witnessed a rapid development of markerless motion capture and understanding for human daily activities, which benefits various real-world applications such as immersive VR/AR experience, action quality assessment (Pan et al., 2019) and vision-based robotics (Ran et al., 2017). How to further capture professional non-daily human motions and provide fine-grained analysis has recently received substantive attention.

In this paper, we focus on markerless motion capture and fine-grained understanding for challenging professional human movements which are essential for many applications such as training and evaluation for gymnastics, sports, and dancing. However, these professional movements like diving and balance-beam suffer from complex motion patterns and severe self-occlusion, especially under the monocular setting, leading to inferior results and impractical usage of existing 3D motion capture (Xiao et al., 2018; Kocabas et al., 2020) and 2D pose detection approaches (Cao et al., 2019). When motion capture is unreliable, further motion

analysis is even more challenging, which aims to provide both mid-level sub-motion categories and detailed semantic descriptions for each sub-motion or the whole motion sequence at the finest granularity. On the other hand, even though it's natural to split such challenging sports movements into sub-motions due to their repeatability and self-similarity, the literature on utilizing such sub-motion prior to strengthening the motion capture and understanding is sparse. Moreover, most existing action understanding solutions (Parmar and Morris, 2019b; Shao et al., 2020) are limited to the pure high-level action assessment, where the abundant 3D motion capture information of sub-motions has been ignored.

To tackle these challenges, we propose *SportsCap* – the first joint 3D motion capture and fine-grained understanding approach for various challenging sports movements from only a single RGB video input (see Fig. 1 for an overview). With the aid of mid-level sub-motion embedding analysis of plausible motion manifold, SportsCap explores and validates the mutual gain between 3D motion capture and fine-grained motion understanding. Our novel pipeline not only achieves significant superiority to previous capture methods for challenging motions, but also provides accurate fine-grained semantic assessment simultaneously for motion understanding, whilst still maintaining a monocular setup.

More specifically, we formulate this joint human motion capture and understanding problem in a multi-task learning framework. To this end, we first introduce a motion embedding space to model the manifold of plausible human poses for each sub-motion via the principal component analysis (PCA) technique. Then, for the motion capture task, an effective motion embedding network is proposed to estimate the per-frame implicit embedding parameters so as to recover the 3D motion details via a corresponding mapping function as well as a sub-motion classifier. Our motion capture scheme leverages the rich semantic and temporally structural prior of sub-motions in the motion embedding space to tackle the severe occlusion and depth ambiguities inherent to the monocular sports video input. For further motion understanding task, we predict the fine-grained semantic action attributes using a spatial-temporal Graph Convolutional Network(ST-GCN), based on both the original motion embedding stream and the recovered 3D detailed motion stream of the whole video clip. Our novel multi-stream ST-GCN module encodes both the implicit and explicit motion information from the previous capture stage for more accurate action attribute parsing. Finally, a semantic attribute mapping block is adopted to assemble various correlated action attributes into a high-level action label for the whole sequence (i.e. the diving number for the diving motion), which provides an extra overall detailed understanding of the whole video to enable various applications like action assessment or motion scoring. To summarize, the main contributions of SportsCap include:

- We propose a novel join human 3D motion capture and motion understanding scheme in a data-driven multi-task manner under the monocular setting, achieving significant superiority to existing state-of-the-arts.
- By utilizing the semantic and temporally structured motion prior in the embedding space, we propose a novel motion embedding module, as well as an effective multi-stream ST-GCN module to reconstruct both detailed 3D motions and accurate fine-grained actions, attributes simultaneously.
- We make available our Sports Motion and Recognition Tasks (SMART) dataset, consisting of various challenging sports video clips with manually annotated poses and fine-grain action labels as well as the relevant ground truth 3D poses for motion embedding analysis.

2 Related Work

Our work is closely related to pose/shape estimation and action parsing. We will discuss related methods and datasets in these research areas.

Pose and Shape Estimation aims to recover the underlying kinematic structure of a person. The results of these work can be 2D/3D poses or 3D human models that match the image/video observations. Earlier methods adopted geometric constraints (Yang and Ramanan, 2011) to construct poses. Recently, with the success of deep neural networks in many computer vision tasks, many deep learning-based pose estimation approaches (Cao et al., 2019; Pishchulin et al., 2016; Raaj et al., 2019; Sun et al., 2019, 2018; Tang et al., 2018) have achieved remarkable performance. Openpose (Cao et al., 2019) employs Part Affinity Fields (PAF) to support bottom-up estimation. Sun et al. (2019) exploits multi-scale high-resolution networks to improve feature representation. However, such methods focus on regular movements and actions and become limited for professional sports, which consist of more complex poses and occlusions in monocular videos. A few recent approaches aim to tackle special actions. Luvizon et al. (2018) proposes a semantic-based, multi-task learning framework, and Bertasius et al. (2018) tailors a predictor specific to certain actions. Xiaohan Nie et al. (2015) uses a hierarchical structure to decompose an action into sub-poses and further divides each pose into many parts. These approaches do not consider rich semantic information embedded in sports, and the structure constraints within sub-motions. In contrast, our approach explicitly uses the underlying semantic and ordering rules in sport to reduce the complexity of the problem. And we utilize PCA to capture the similarities of poses in each sub-motion and

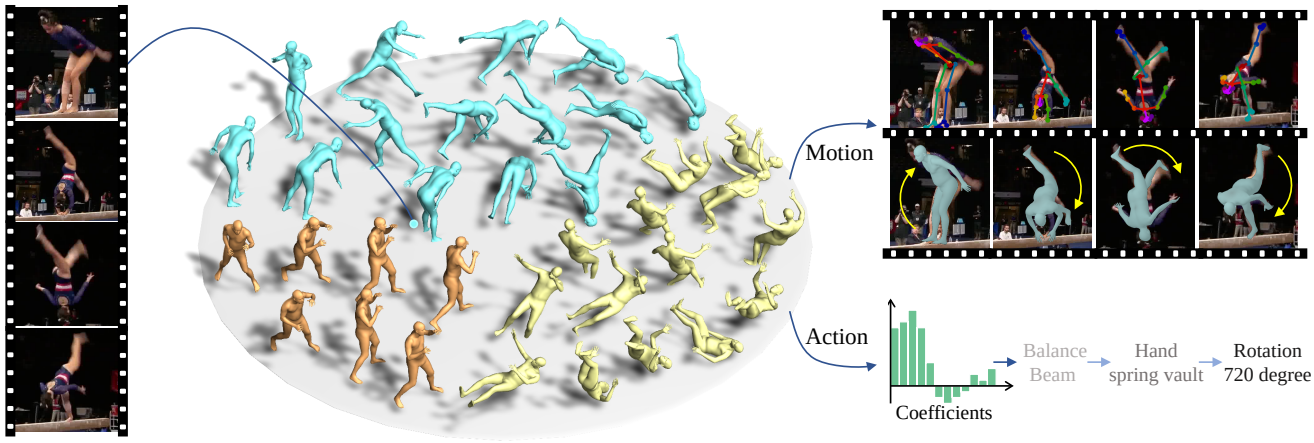


Fig. 1 SportsCap: A multi-task approach for 3D motion capture and action understanding of challenging sports videos. We collect a sport-related motion capture dataset to build the Sports Pose Embedding Spaces on specific sports, like balance beam (Blue), boxing (Orange), high jump (Yellow), and other sports. This Sports Pose Embedding Spaces achieve significant superiority on challenging motion capture and encode semantic meanings (see Fig. 5) for action parsing tasks.

constrain estimated poses in reasonable forms to further improve the accuracy.

Traditional 3D human estimation methods either use multi-camera dome systems (Kanade et al., 1997; Collet et al., 2015) or exploit the RGB-D sensors (Dou et al., 2016; Newcombe et al., 2015), and recover the human geometry via multi-view stereo and point cloud fusion. With the advance in parametric 3D human body models and deep neural networks, especially with the emergence of the SCAPE (Anguelov et al., 2005), SMPL (Loper et al., 2015), SMPL-X (Pavlakos et al., 2019), recovering the human shape from a single viewpoint image/video becomes more and more popular. SMPLify-X (Pavlakos et al., 2019) fits the face, hand, and body parts of the SMPL-X model to images with pre-estimated 2D poses. HMR (Kanazawa et al., 2018a) proposes an end-to-end framework to regress the pose/shape parameters of human model directly from a single image supervised by an adversarial prior. Similarly, the VIBE (Kocabas et al., 2020) leverages the human pose data from the motion capture dataset (AMASS (Mahmood et al., 2019)) and develops an adversarial framework to discriminate between real human motions and the produced temporal pose and shape. Recovering the human shape from competitive sports images/videos is even more challenging. The aforementioned methods rely on 2D poses to regress the human pose and shape parameters, while athletes in competitive sports exhibit highly complex poses and fast motions that won't appear in daily activities. We tackle the problem through embedding the highly complex but standard human poses in sports to parametric space. Furthermore, we collect the pose data for competitive sports activities using a MoCap system to provide more reliable priors and use these priors to constrain the estimated poses in reasonable forms.

Action Parsing can be categorized into short vs. long dynamics, depending on the length of the motion patterns. For short term dynamics, Karpathy et al. (2014) uses 2D CNNs to learn deep appearance features and conduct frame-level classification. IDT (Feichtenhofer et al., 2016) extends the technique with shallow motion features and Hussein et al. (2019) uses 3D CNNs such as C3D (Tran et al., 2015) to capture spatial-temporal patterns of consecutive frames within the sequence. For long term dynamics, TRN (Zhou et al., 2018) exploits temporal dependencies across video frames over multiple hierarchies. TRN (Zhou et al., 2018) proposes a multi-stream architecture to extract even richer temporal features. LTC (Varol et al., 2017) treats the temporal resolutions as a substitute to temporal windows whereas Hussein et al. (2019) conducts long-range action recognition. We observe that competitive sports, like diving and gym, are always a mixture of long and short dynamics: sub-actions such as twisting or somersaults map to short dynamics whereas the complete dive, with a corresponding dive number, map to long dynamics. We hence combine the advantages of short and long dynamics techniques.

To specifically tackle sports videos, Kanojia et al. (2019) proposes an attentive guided LSTM-based neural network for fine-grained motion recognition. Pishchulin et al. (2014) combines the dense motion trajectories and pose estimations to improve recognition accuracy. Choutas et al. (2018) represents the movement of semantic keypoints as a color encoded trajectory map, called PoTion, and subsequently conducts classification on the PoTion. In a similar vein, Fani et al. (2017) stacks the poses features generated by an hourglass network into a reference frame and then performs the sub-action recognition from hockey sports videos. Pan et al. (2019) builds a joint relation graph to model both the joint relations within a time step and across two immediate time steps. Nevertheless, though these approaches rely on the joint

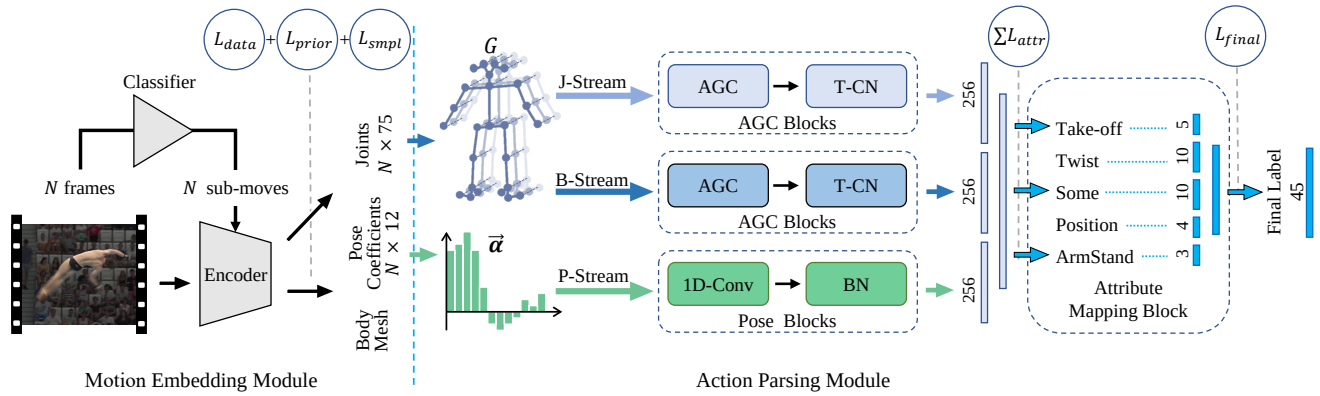


Fig. 2 Our SportsCap is composed of two main components: the Motion Embedding Module and the Action Parsing Module. Motion Embedding Module estimates motion embedding information, 3D joints, and 3D body meshes, while Action Parsing Module predicts the fine-grained semantic attributes and final action labels of sports.

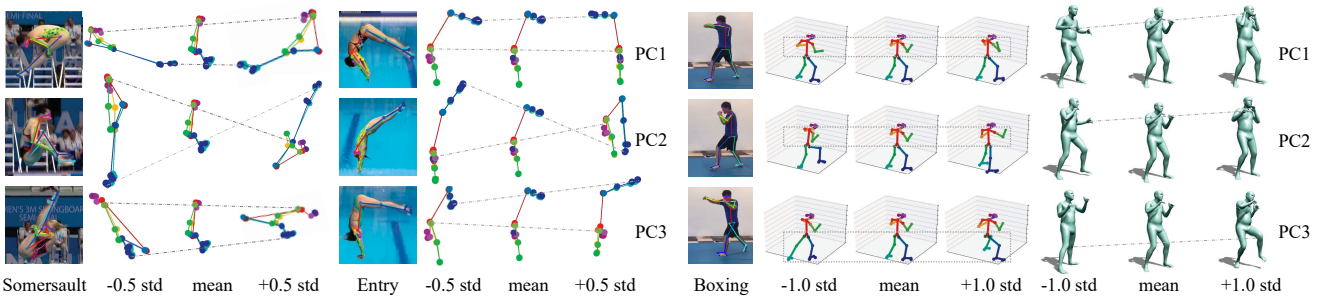
motions for action recognition, they ignore the patterns of the human body motion in certain activities. In contrast, we observe that the joint motion within a sub-action tends to be regular in competitive sports. Hence, we adopt a fine-grained manner to model the pose in each sub-action and finally resorts to the recent Graph Convolutional Network (GCN) (Defferrard et al., 2016; Henaff et al., 2015; Li et al., 2018a,b; Shi et al., 2019) for spatial-temporal representations.

Dataset is the basis for deep learning-based motion estimation and action parsing methods. There are some large-scale human image/video datasets, such as COCO (Lin et al., 2014) and MPII (Andriluka et al., 2014a). They mainly focus on motions in daily motions. Competitive sports video understanding relies heavily on available sports datasets. Zhang et al. (2013) proposes a simple motion dataset of 15 actions with annotated body joints but no action labels. Parmar and Morris (2019a,b) presents the MTL-AQA dataset that exploits multi-task networks along with a caption generation model to simultaneously assess the move and produce a caption. Li et al. (2018c) proposes the Diving48 dataset for competitive diving video understanding. The UCF101 dataset (Soomro et al., 2012) contains 101 classes of in-the-wild actions and the ActivityNet Caba Heilbron et al. (2015) covers a wide range of complex human activities in daily living. More recently, Shao et al. (2020) proposes the FineGym dataset which contains 10 event categories, 303 competition records and provides coarse-to-fine annotations both temporally and semantically. Competitive sports videos contain both rich semantic action information and strict human body motions. Similar to the FineGym and Diving48, our SMART dataset contains per-frame annotated action labels. In addition to the fine-grained semantic labels, we further add manually annotated human pose, MoCap pose space of each sub-action, and action assessment from professional referees. To our knowledge, the SMART dataset is the only one that provides the fine-grained semantic labels, 2D and 3D annotated poses, and assessment information.

3 Overview

Our goal in this paper is to reconstruct both the 3D human motion and the corresponding fine-grained action attributes from monocular professional sports video input. To handle this challenging problem, our SportsCap splits each professional motion into a sequence of elementary sub-motions, and utilizes the motion manifold prior of these sub-motions in a multi-task learning framework, as illustrated in Fig. 1. Our approach not only captures the fine 3D motion details for each sub-motion, but also provides detailed motion understanding attributes, such as “take-off type” for the “take-off” sub-motion and “somersault number” for the whole sequence (this is because different “take off” type and “entry” type will also contribute to “somersault number” attribute). To model this motion capture and understanding problem in a data-driven manner, we collect a new Sports Motion and Recognition Tasks (SMART) dataset, which contains various challenging sports video clips with manually annotated ground truth poses and fine-grain action labels as well as the corresponding relevant 3D poses captured via a motion capture system. A brief introduction of the two main components of our pipeline is provided as follows, which explores and proves the mutual gain between 3D motion capture and fine-grained motion understanding.

Motion Embedding Module. For a challenging sports video, we propose a motion embedding space to model the manifold of plausible human poses for each sub-motion via the PCA technique. Based on such embedding prior, we further introduce a novel network to estimate the per-frame implicit motion embedding parameters so as to recover the 3D motion details, including the pose, shape parameters of the statistical human model SMPL Loper et al. (2015) and cam-



(a) Diving Pose Embedding Space (b) Boxing Pose Embedding Space

Fig. 3 Motion embedding pose spaces of 2D poses, 3D poses and pose parameters of SMPL: (a) 2D Pose embedding spaces for somersault and entry (two sub-motions of competitive diving). (b) Motion embedding spaces of 3D poses and pose parameters of SMPL for boxing. For each sub-motion, from left to right, we show the input frame and the first three principal components(PC) within 0.5 or 1 deviations from the mean. The lines connecting the corresponding elements within the component indicate the linear change according to the basis.

era parameters. Our embedding module consists of a sub-motion classifier, a CNN encoder to regress the embedding, and the corresponding mapping function from the embedding space to the 3D motion output (see Sec. 4.1).

Action Parsing Module. For further motion understanding tasks, we propose to predict the fine-grained action attributes for the whole motion sequence using a novel multi-stream ST-GCN module, which makes full use of both the implicit pose embedding and explicit 3D joints and bones from the previous capture stage. We further propose a semantic attribute mapping block to map the predicted attributes to the final action label, which enables various applications such as action number prediction (like diving or gymnastics number) for motion scoring and action assessment.

4 Technical Details of SportsCap

Fig. 2 shows the architecture of our SportsCap, which takes a sports video as input and reconstruct both 3D motion details and accurate action attributes in an end-to-end multi-task manner. We assume the input video clip corresponds to the complete sports motion and split it into several sub-motions, which are segments that correspond to its sports stages similar to previous work (Hu and Qi, 2019). For example, the four stages of diving are take-off, twist, somersault, and entry. Our SportsCap consists of two modules for both per-frame 3D pose/shape reconstruction and action understanding such as per sub-action labeling and action assessment. The recovered pose and shape parameters can be further applied to drive a 3D parametric human model to conduct the same sport movement in 3D. For motion capturing, we construct their corresponding motion embedding functions for respective segments where each frame is fed to its respective encoder to obtain its motion embedding information, joints, and bones. For action labeling, we construct a multi-stream ST-GCN that takes the coefficients, joints, and bones as input and conduct multi-task action attribution prediction. Our ST-GCN contains an attributes mapping block that assembles action attributes into the final label, which indicates an action number or score.

4.1 Motion Embedding and Capturing

Professional poses in sports have complex structure information and always bring occlusions in monocular videos, which impose significant challenges to existing pose/shape estimators such as OpenPose (Cao et al., 2019), Simple-Baseline (Xiao et al., 2018) HMR (Kanazawa et al., 2018a) and VIBE (Kocabas et al., 2020). Fig. 11 shows some typical results. This is partially due to the pose variants in sports. More importantly, those approaches do not explore the specific semantic and structural constraints in sports.

We present a novel Motion Embedding Module that takes into account the structure of pose spaces of a specific sport. We recognize that the complete sport move in profession always follows several stages, called sub-motions. A sub-motion is a segmentation of the video sequence in the temporal domain according to movement regularity, which does not have to be semantically meaningful. For example, boxing action can be segmented into three sub-motions: punching, kicking and dodging. And the diving action has four sub-motions, i.e. the take-off, twisting, somersault, and entry. In each sub-move, the poses exhibit high resemblance across athletes, e.g., divers straighten their bodies in twisting while curling up in somersault. Considering these characteristics in sports, we construct a motion embedding function for each sub-move to capture the structural similarities.

Motion Embedding Module aims to build a parametric pose space with specific constraints for each sub-motion. We represent the pose parameters of SMPL as $\boldsymbol{\theta}$. For a sub-motion m, the motion embedding function is formulated as

follows:

$$\theta = \mathcal{M}_m(\boldsymbol{\alpha}), \tag{1}$$

$$\theta = \sum_{k=1}^{K} \alpha_k \mathbf{b}_k^m + \mathbf{a}^m = \boldsymbol{\alpha}^\intercal \mathbf{B}^m + \mathbf{a}^m, \tag{2}$$

where $\mathcal{M}_m(\boldsymbol{\alpha}) : \mathbb{R}^K \mapsto \mathbb{R}^{3N}$, N denotes the joint number of SMPL, and K denotes the dimension of pose coefficients $\boldsymbol{\alpha}$. $\mathbf{a}^m$ is the mean of pose parameters and $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_K]^\intercal$ are the pose coefficients. Although the poses of the sports are challenging, the similarity of the poses in a sub-motion shows a desirable feature on lower-dimension. Thus, we adopt the Principal Component Analysis (PCA) to model the pose space of each sub-motion. We collect a mocap dataset with more than 50 thousand frames to provide the set of pose parameters $\{\boldsymbol{\theta}^i\}$ and conduct PCA on $\{\boldsymbol{\theta}^i\}$ to generate a set of pose bases $\mathbf{B}^m = \{\mathbf{b}_k^m\}_{k=1}^K$, so that θ under sub-motion m can be represented as a linear combination of the bases. With our mocap data, we calculate all pose bases $\{\mathbf{B}^m, \mathbf{a}^m\}$ for each sub-motion m before our training parts.

We not only formulate the motion embedding function (Eq. 2) on pose parameters, but also formulate it on 2D/3D joints as $\mathbf{J} = \mathcal{M}'_m(\boldsymbol{\alpha})$. Fig. 2 visualizes the motion embedding spaces on 2D joints, 3D joints, and pose parameters of SMPL, which describes the first three pose bases in somersault, entry (two sub-motions of dive), and boxing. The approach reduces the dimension of poses, benefiting for training, and regression. It can also robustly handle all sub-motions even for traditionally challenging poses because of the extracted structure of pose spaces. The resulting pose coefficients representations also encode semantic meanings (see Fig. 5), like rotation angle, important for subsequent action parsing Sec. 4.2.

From the pose bases for all sub-motions, we construct the Motion Embedding Module that estimates the pose coefficients and 3D joints. This motion embedding representation can be suitable for many kinds of backbones. In our case, Motion Embedding Module consists of a ResNet-152 convolutional encoder followed by 2 fully connected layers to regress the pose coefficients $\boldsymbol{\alpha}$ used to reconstruct the joint positions:

$$\boldsymbol{\alpha}(\mathbf{x}) = \mathcal{F}_{conv}^m(\mathbf{x}; \mathbf{W}), \tag{3}$$

$$\mathbf{J}(\mathbf{x}) = \boldsymbol{\alpha}(\mathbf{x})^\intercal \mathbf{B}^m + \mathbf{a}^m, \tag{4}$$

where $\mathbf{x}$ denotes input image/frame and $\mathcal{F}_{conv}^m$ is the motion embedding network for the sub-motion m. of shape capturing. We utilize this network to estimate pose coefficients, shape parameters and camera parameters (see Eq. 6) from images. Then, we recover 3D human body meshes from estimated pose and shape parameters With SMPL, the parametric human model.

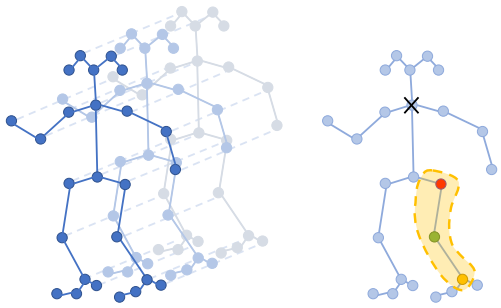


Fig. 4 We show our spatial-temporal graph of the joints and bones. Right shows the spatial configuration partitioning strategy: we divide the node's one neighbor into three subsets: the root node (green dot), the centripetal subset (red dots), and the centrifugation subset (yellow dots), details in Li et al. (2018a).

Unlike prior approaches that target at general poses by implicitly encoding pose regularity into a complex network and hence cannot reliably handle motion blurs and occlusions or easily enforce semantic constraints, Motion Embedding Module manages to exploit the structural constraints in sport poses with action semantics. Even though we formulate the motion embedding equation (Eq. 4.1) with the pose parameters of SMPL, Motion Embedding Module can also be applied to other joint representations, like 2D/3D joint location/rotation, which is used in our experiment of 2D pose estimation (see Fig. 11).

4.2 Action Parsing

Next, we feed the estimated pose coefficients and 3D joint positions of all frames from Motion Embedding Module to the Action Parsing Module for analyzing the complete action, including inferring semantic attributes and the action number from the sequence and later assessing the performance. Semantic Attributes (SAs) represents the semantic meaningful label assumed by the sport/exercise guidelines, rules, or people. For example, the five SAs for a diving action are the take-off type, twisting number, somersault number, arm-stand, and dive position (see Fig. 7(b)). The action number represents a valid combination of SAs, which is an overall description of a sports action. For example, certain diving actions may only contain a subset of these SAs. (see Fig. 7(b)) The brute-force approach would be to build a black-box network to map the pose sequence to action number.Competitive sports have detailed defined elements and fixed semantic attribute types. So, we adopt a different approach that treats the action number as attributes of the action. Recall the action number encodes key semantic meaning of the sport move, e.g., the number of rounds of somersault in gymnastics or dive, the number of punching in boxing, etc. (see Fig. 7(b) for more details). We call them semantic attributes (SAs) and aim to learn how each

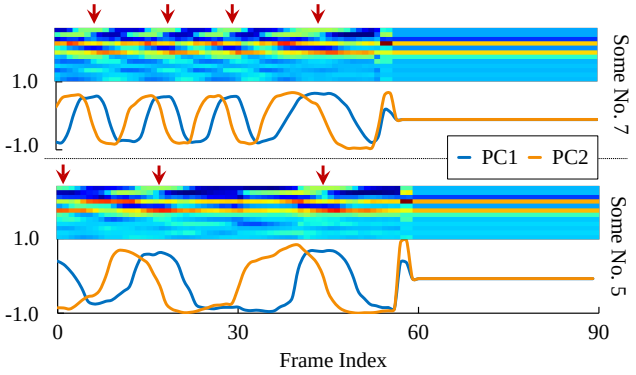


Fig. 5 Extracted feature maps of pose coefficients from the P-stream encode the semantic meaning of sport. Each feature map is of 90×25 (90 frames and 25-dimensional features)In each figure, the upper half is the principal components feature map of the entire sequence, and the lower half is the specific numerical curve of the first two principal components.. The first two principal components already reveal the number of rounds of half-somersault (left: $4 \times 2 - 1 = 7$ rounds; right: $3 \times 2 - 1 = 5$ rounds). Red arrows correspond to the start position of a somersault (toe pointing to the ground) whereas the final half-somersault corresponds to entry to the water.

frame contributes to respective attributes. Our Action Parsing Module explicitly recovers SAs via a two-stage architecture: in the first, we use a multi-Stream ST-GCN for SA predictions, and in the second, a attributes mapping block infers the action number or score from the SAs.

Spatial-Temporal Feature Extraction. A number of previous approaches such as (Wen et al., 2019; Si et al., 2018; Zhang et al., 2019) exploit skeletons alone as inputs to the GCN. our pose coefficients encode meaningful semantic submotion using proposed motion embedding analysis, in addition to skeletons (bones and joints), pose coefficients obtained from Motion Embedding Module provide useful information on action parsing, as shown in Tab. 4. We therefore construct a multi-Stream convolutional module that takes joints (J-stream), bones (B-stream), and pose coefficients (P-stream). For J- and B-Stream, we adopt the 2s-AGCN structure that can adaptively learn graph edge connections. Details on graph construction and partitioning are shown in Fig. 4.1. Specifically, we adopt the human joints and bones setting in OpenPose(Cao et al., 2019). In the J-stream, the joints are mapped to graph nodes, and the bones map to edges. In the B-Stream, the mappings are reversed. We feed 90 consecutive frames of skeletons into a 10-layer ST-GCN to extract two feature vectors. For the P-stream, we represent pose coefficients as a 1D vector and use layers of 1D convolution with residual blocks to generate features of 256 dimensions. We demonstrate the effectiveness of the proposed P-Stream. In Fig. 5, we visualize one of the feature maps of a specific sequence obtained from the P-Stream, at a resolution of 90×25 (90 frames and 25 dimensions in feature). We also plot the first two dimensions vs. frame index. We observe that they can be readily used to infer the somersault number of competitive diving. We finally concatenate all feature vectors generated by the J-, B- and P-Stream as inputs to the following attributes mapping block.

Semantic Attributes Mapping Block. The Semantic Attributes Mapping Block aims to learn the mapping between the extracted spatial-temporal features and the final action label, i.e., to tell which dive number or action scores the video corresponds to. Instead of directly learning the mapping via a black-box solution, we sought to use Semantic Attributes (SAs) explicitly. Specifically, our goal is to partition the whole action sequence in terms of the SAs, or more precisely, how individual frames contribute to specific SAs. We use two fully connected layers to predict their contributions where the categories of all SAs are represented as vectors. Finally, we stack the resulting SAs and feed the results to another two fully connected layers to infer the action number. Compared with black-box end-to-end approaches, our results show the use of intermediate SAs better supervise the training process, provide heuristic cues analogous to human labeling, and accelerate the training process. Moreover, decomposing the whole action into the SAs resembles human perception, which helps to analyze the fine-grained action of sports videos.

4.3 Multi-task Training

To train our end-to-end multi-task network, we adopt a deeply-supervised strategy in which we design five losses for Motion Embedding Module and Action Parsing Module.

Loss for Motion Embedding Module. In our network, we use three types of representations, pose coefficients, pose parameters and 3D joints, to model poses. The pose coefficients define the motion embedding space of the pose, whereas the joint positions better describe the visibility between joints within an image. We therefore design the prior loss (the pose coefficients loss) $\mathcal{L}_{prior}$ as:

$$\mathcal{L}_{prior} = \left\| \mathbf{W}(\overline{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}) \right\|_2 , \tag{5}$$

where $\overline{\boldsymbol{\alpha}}$ is the mean of pose coefficients in training set. $\hat{\boldsymbol{\alpha}}$ is the predicted pose coefficients. Then, $\mathbf{W}$ is the weights, which calculates from the eigenvalues of a covariance matrix in the calculation of pose embedding bases. We use the smaller weight for the larger eigenvalue to enable more tolerance on principal components. We design the data loss (the 2D joint position loss) $\mathcal{L}_{data}$ as:

$$\hat{\mathbf{J}} = \hat{s}\Pi(\mathcal{J}(\mathcal{M}_m(\hat{\boldsymbol{\alpha}}), \hat{\boldsymbol{\beta}})) + \hat{t}, \tag{6}$$

$$\mathcal{L}_{data} = \left\| \mathbf{V}(\mathbf{J} - \hat{\mathbf{J}}) \right\|_2 , \tag{7}$$

where $\mathbf{J}$ is the ground truth 2D joints and $\mathbf{V}$ indicates the visibility of the ground truth joint. Pose parameters $\hat{\boldsymbol{\theta}} = \mathcal{M}_m(\hat{\boldsymbol{\alpha}})$ is recovered from our embedded pose coefficients

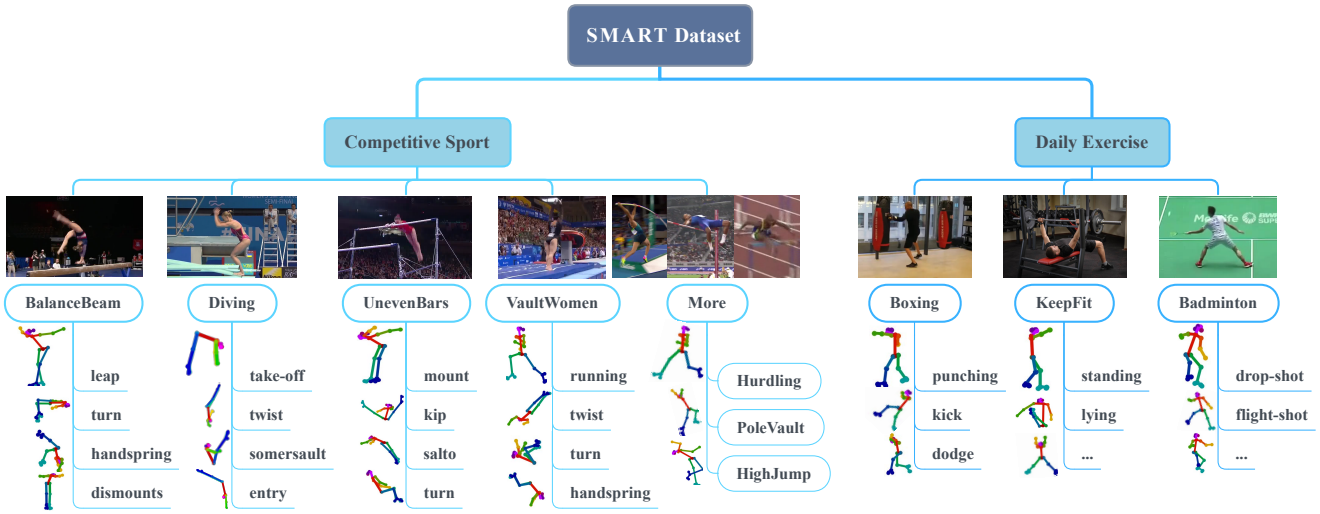


Fig. 6 Our SMART dataset contains both pose and action annotation. The upper part shows specific categorices of sport. The lower part depicts the sub-motions of each sport and its typical poses.

$\hat{\boldsymbol{\alpha}}$, using this motion embedding function of the sub-motion m. 3D joints $\mathcal{J}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}})$ are obtained by linear regression from the final mesh vertices of SMPL. We then follow Kanazawa et al. (2018a) using a weak-perspective project system with only scale s, translation parameters $t, t \in \mathbb{R}^2$, and the orthographic projection function $\Pi(\cdot)$. We design the SMPL loss as:

$$\mathcal{L}_{smpl} = \left\| \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right\|_2 + \left\| \boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \right\|_2, \tag{8}$$

where $\boldsymbol{\theta}$,$\boldsymbol{\beta}$ are the supervision of pose/shape parameters, which are obtained through MoSh (Loper et al., 2014) and provided mocap data.

We combine these three loss terms, the prior loss of our motion embedding, the data loss, and the pose/shape parameter loss of SMPL with the corresponding weights λ_{data}, λ_{smpl} (10 and 2 in our case) as the final loss of the Motion Embedding Module:

$$\mathcal{L}_{mem} = \mathcal{L}_{prior} + \lambda_{data}\mathcal{L}_{data} + \lambda_{smpl}\mathcal{L}_{smpl}. \tag{9}$$

Loss for Action Parsing Module. For the Action Parsing Module, we use the cross-entropy loss between the predicted and the ground truth attributes, which can be written as follows,

$$\mathcal{L}_{attr} = \sum_{c=1}^{Ns} \sum_{i=1}^{N_c} y_i^c \log\left(x_i^c\right), \tag{10}$$

where c indicates the attribute type, N_s denotes the number of attributes, and N_c is the categories of each attribute c.For example, the number of somersault in diving has 10 categories, indicating the rotation angle for 0 to 3240 ($0 * 360$ to $9 * 360$) degrees. Here x_i^c denotes the prediction for the i-th label of attribute c whereas y_i^c is the ground-truth.

For the action labeling task, we also add the cross-entropy loss between the prediction and the ground truth action label as the task loss as below. We note that such a task loss depends on the target application and can be easily adjusted according to the final task.

$$\mathcal{L}_{task} = \sum_{j=1}^{N_f} y_j \log\left(x_j\right), \tag{11}$$

where N_f denotes the total number of all possible action labels (action number or score), and x_j, y_j are the prediction and ground truth of the j-th final label.

The overall loss for the Action Parsing Module is a combination of the attribute loss and the task loss as following:

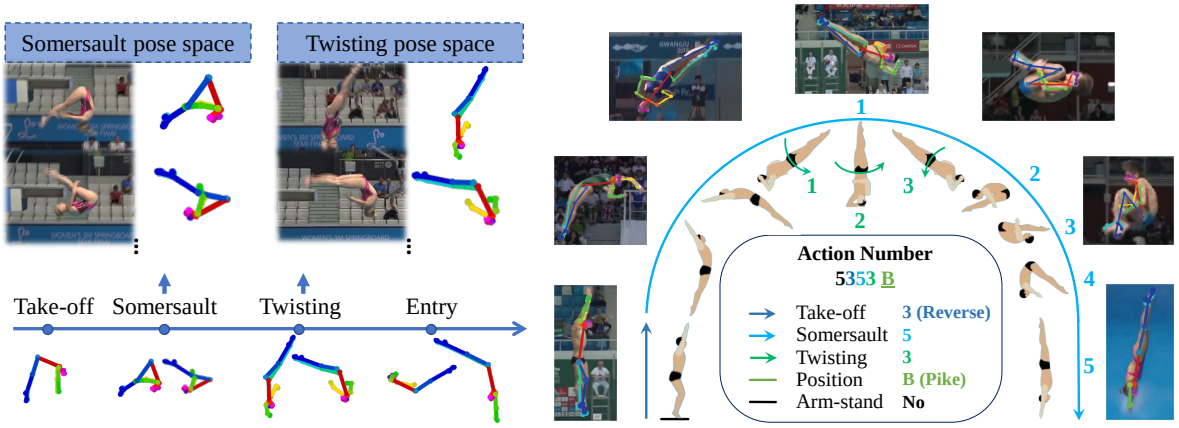
$$\mathcal{L}_{apm} = \mathcal{L}_{attr} + \lambda_A \mathcal{L}_{task}, \tag{12}$$

where λ_A (2 in our case) is a weight to balance two loss terms.

Training Strategy. While our entire network can be trained in an end-to-end fashion, we exploit its modular architecture and develop a stage-wise strategy, which is more efficient in practice. Specifically, our training procedure is composed of three stages: 1) We train a Motion Embedding Module for each sub-motion independently and fix its parameters, 2) We then train the action attribute prediction and label classification modules in the Action Parsing Module jointly, and 3) Finally we fine-tune the entire network using the combined losses of Motion Embedding Module and Action Parsing Module (i.e., $\mathcal{L}_{mem} + \mathcal{L}_{apm}$).

5 Experimental Results

In this section, we evaluate our SportsCap in a variety of challenging scenarios. We first report a new proposed dataset



(a) Sub-motions of competitive diving. (b) Semantic attributes of competitive diving.

Fig. 7 Two important types of label annotations in the SMART dataset, sub-motions (SMs) and semantic attributes (SAs). We introduce SAs and SMs of competitive diving as example. (a) SMs indicate the different pose spaces, usually different stages in a sport, like somersault and twisting in this case. (b) SAs indicate the specific number of a motion, like the rotation angle, take-off type and so on.

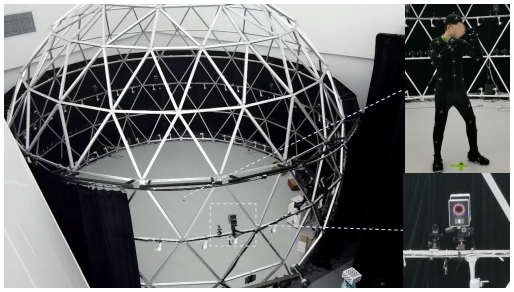


Fig. 8 The motion capture system. The structure of dome is shown in the left and the motion capture camera is shown in the bottom right. Our system includes 12 multi-view motion capture cameras. The top right figure shows the motion cap suit with 63 marks. With professional performers, we utilize this system to capture 3D challenging motion of sports.

and the training details of our approach on the utilized datasets, followed by the evaluation of our main technical components, including both qualitative and quantitative comparison with previous state-of-the-art methods on both motion capture and action parsing tasks. Finally, limitations and discussions regarding our approach are provided.

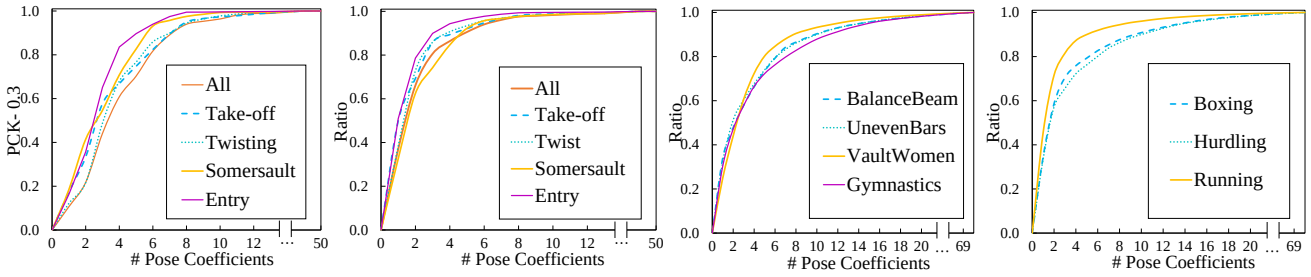
5.1 Dataset

Many datasets exist for human action analysis, such as the MPII (Andriluka et al., 2014a) and COCO (Lin et al., 2014) for human pose in daily activities, the PennAction (Zhang et al., 2013) for coarse sport recognition, and the AQA (Parmar and Morris, 2019b) and FineGym (Shao et al., 2020) dataset for fine-grained action recognition. We propose a challenging sports dataset called Sports Motion and Recognition Tasks (SMART) dataset, which contains per-frame action labels, manually annotated pose and action assessment of various challenging sports video clips from professional referees. We also collect the human pose data in sports activities using a marker-based motion capture system (Fig. 5) to provide pose prior to our Motion Embedding Module. To our best knowledge, our SMART dataset is the most complete dataset for human motion capture and action analysis for sport video (see Tab. 1).

Our SMART dataset consists of both competitive sports and daily exercise videos, including the high jump, competitive diving, uneven bars, balance beam and vault-women in competitive sports, and running, badminton, boxing, and keep-fit in the daily exercise category. The SMART has 640 videos (110K frames) in total, with per-frame skeleton annotation and sub-motion labels, semantic attribute labels, and action assessment scores for competitive sports. There are about 450,000 annotated skeletons (in the Openpose 25 joints representation (Cao et al., 2019)) with corresponding bounding boxes. In addition to joint locations, we also annotate the visibility of each joint as three types: visible, labeled but not visible, and not labeled (same as the COCO (Lin et al., 2014)). To fulfill our goal of 3D pose estimation and fine-grained action recognition, we collect two types of annotations, i.e. the sub-motions (SMs) and semantic attributes (SAs), as we described in Sec. 4.1 and Sec. 4.2. We also include the difficulty scores, the number of valid referees, the execution scores, and the final assessment scores for competitive sports in the SMART dataset. We will share our SMART dataset with the community.

In addition to the annotated video data, we also collect the 3D human pose data using a motion capture system to provide pose motion prior for our Motion Embedding Module. Specifically, we adopt the Vicon system (a 12 views marker-based motion capture system) to capture the rich human pose sequences and hire 30 athletes and 2 professional

Fig. 9 3D human shape recovery results on challenging sport videos. For each type of sport, the top row shows the input images, the middle row shows the recovered body mesh, and the bottom row shows the rendering result of the recovered body from an alternative view.



(a) Diving - PCK-0.3 vs. PCA dim(dimension). (b) Diving - The ratio vs. PCA dim. (c) Gymnastics - The ratio vs. PCA dim. (d) Other sports - The ratio vs. PCA dim.

Fig. 10 Cumulative relative variance of our sports dataset explained as a function of the number of pose coefficients. (a) and (b) are the sub-motions of diving with different metrics. 'All' refers to use all training data rather than restricted to each semantic pose space. (c) and (d) are general motion embedding of gymnastics and daily sport. PCKs indicates the percentage of correct keypoints (see details in (Andriluka et al., 2014b)), and the ratio follows (Loper et al., 2015).

Table 1 Comparison between the SMART dataset and existing datasets, including MPII-Pose (Andriluka et al., 2014a), Penn Action (Zhang et al., 2013), COCO (Lin et al., 2014), AQA (Parmar and Morris, 2019b) and FineGym (Shao et al., 2020), regarding size, per video action labels, pose annotation, action assessment, and pose type. We show the distribution and annotation details of each specific sport of SMART in the bottom of the table.

Dataset	Video Clips	Images	Actions per clip	Pose	3D mocap	Assess	Type
MPII	-	25K	-	✓	×	×	General
COCO	-	330K	-	✓	×	×	General
Penn	2K	160K	Single	✓	×	×	Sport
AQA	1K	150K	Multi	×	×	✓	Dive
FineGym	5K	1.9M	Multi	×	✓	×	Gymnastics
Ours	**5K**	**2.1M**	**Multi**	✓	✓	✓	**Sport**
-	640	110K	Multi	✓	15K	✓	Dive
-	2K	434K	Multi	✓	5K	✓	Vault-Women
-	1K	490K	Multi	✓	5K	✓	UnevenBars
-	1K	704K	Multi	✓	2K	✓	BalanceBeam
-	-	24K	Single	✓	5K	×	Hurdling
-	-	155K	Single	✓	2K	×	PoleVault
-	-	96K	Single	✓	3K	×	HighJump
-	-	4K	Single	✓	11K	×	Boxing
-	-	54K	Single	✓	2K	×	Badminton

fitness instructors as our performers. The performers move according to the guidelines of the corresponding sports to make sure their body movements cover as more sub-motions of the sports as possible (except the motions that are impossible to execute in the capture environment). For the generalization purpose, each performer repeats the movements two to five times and the same motion is captured from more than two different subjects. Since only the relative motion matters, we convert the skeleton results from Vicon to the SMPL pose parameters to avoid the variations imposed by the absolute lengths of bones. In total, we collect more than 500,000 motion frames of 30 performers covering 9 activities, about 1,000 frames for each performance.

With the annotated 2D poses and MoCap 3D pose data, we collect the Sports Pose Embedding Spaces according to our motion embedding function (Eq. 4.1) and use it as the pose priors for sports videos. Currently, Sports Pose Embedding Spaces provides the priors on 2D joints, 3D joints location, and pose parameters of SMPL for 9 sports, as shown in Tab. 1, and we are planning to add more sports in the near future. Because of the regularity of the human body motion in sports/exercises, the Sports Pose Embedding Spaces provides strong prior and regularization to ensure that the generated pose result lies in the corresponding action space. The Sports Pose Embedding Spaces greatly improves the accuracy and robustness of the 2D/3D pose estimation, human body capturing, action recognition/parsing, and action assessment tasks as described in Sec. 4.1. The Sports Pose Embedding Spaces data will be included in the SMART dataset.

5.2 Training Details

Our Motion Embedding Module relies on both the fine-grained action labels and pose information, we first train and test the Motion Embedding Module on the SMART dataset as few other datasets provide both information. Then we fix the Motion Embedding Module and train the complete SportsCap on SMART, AQA (Parmar and Morris, 2019b) and FineGym dataset (Shao et al., 2020), for 3D sports motion estimation and action understanding.

We resize image patches that contain the human body at a resolution of 256×256 (using the ground truth bounding box in our SMART dataset and detect the bounding box in AQA (Parmar and Morris, 2019b) using Liu et al. (2016)). We re-sample the video to 90 frames each. For Motion Embedding Module training, we conduct data augmentation via random rotations ($-45°$ to $+45°$), random scaling (0.7 to 1.3), and flipping horizontally. For Action Parsing Module training, we also augment the skeleton and pose parameters data for J-, B- and P-Streams, respectively. For J-Stream and B-Stream, we scale the joint positions via interpolation to

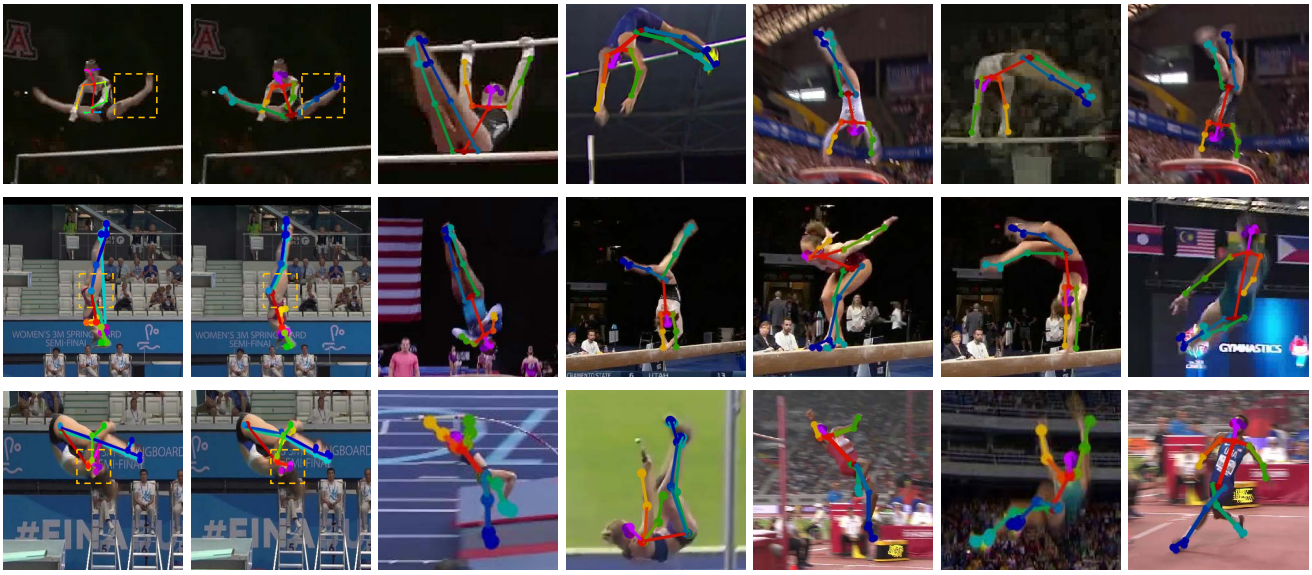


Fig. 11 The gallery and comparison of our experiments on 2D pose estimation. With 2D motion embedding spaces, we show the comparison with fine-tuned HRNet (Sun et al., 2019) (The first column) on SMART dataset. Our results (start from the second column) are more robust and reliable under challenging poses and motion blur.

Table 2 Ablation study (PCA w/o sub-motion labeling, 50/101/152 ResNet as Backbone, and $\mathcal{L}_{\boldsymbol{\alpha}}$ ($\mathcal{L}_{prior}$)/$\mathcal{L}_{\mathbf{J}}$ ($\mathcal{L}_{data}$) as loss for training. With no $\mathcal{L}_{\boldsymbol{\alpha}}$, we follow Kanazawa et al. (2018a) use pose parameters as the variable of $\mathcal{L}_{\mathbf{J}}$ directly.) and comparisons with different methods trained on the same SMART dataset: HRNet (Sun et al., 2019), SimpleBaseline (Xiao et al., 2018), HMR (Kanazawa et al., 2018a), and VIBE (Kocabas et al., 2020) of our Motion Embedding Module. We project the recovered 3D joints from HMR and VIBE into image as their keypoint predictions. SM-0 to SM-3 are the sub-motions of our sports (e.g. take-off, twist, somersault, and entry for diving). We use the percentage of correct keypoints (PCK-0.3, PCK-0.5) as our metrics.

Method	PoseSpace	Backbone	SM-0	SM-1	SM-2	SM-3	PCK-0.3	PCK-0.5
HRNet	-	-	83.9	81.7	82.7	86.4	83.6	87.5
Simple	-	ResNet152	86.3	68.5	86.7	90.4	84.2	88.9
HMR	-	ResNet50	81.3	68.9	71.9	73.1	73.8	84.1
VIBE	-	-	72.0	41.9	78.7	59.4	44.1	62.4
Ours+$\mathcal{L}_{\mathbf{J}}$	-	ResNet50	69.1	60.7	76.5	71.3	70.9	86.5
Ours+$\mathcal{L}_{\boldsymbol{\alpha}}$	Semantic	ResNet50	82.9	75.6	83.1	90.3	83.6	91.5
Ours+$\mathcal{L}_{\boldsymbol{\alpha}}$+$\mathcal{L}_{\mathbf{J}}$	Single	ResNet50	81.0	80.1	82.7	89.8	83.0	92.1
Ours+$\mathcal{L}_{\boldsymbol{\alpha}}$+$\mathcal{L}_{\mathbf{J}}$	Semantic	ResNet50	81.3	80.4	87.6	88.8	84.8	92.4
Ours+$\mathcal{L}_{\boldsymbol{\alpha}}$+$\mathcal{L}_{\mathbf{J}}$	Semantic	ResNet101	**87.2**	82.2	88.6	90.5	87.5	94.5
Ours+$\mathcal{L}_{\boldsymbol{\alpha}}$+$\mathcal{L}_{\mathbf{J}}$	Semantic	ResNet152	83.6	**84.6**	**91.5**	**94.0**	**88.5**	**96.0**

simulate the far and near camera views. In addition, we randomly mirror the skeleton (to emulate viewing from both sides). For the P-Stream, we also scale the coefficients vector.

We use the Adam optimizer (Kingma and Ba, 2015), train the Motion Embedding Module in the first 100 epochs and AMP in the following 50 epochs. We train the complete SportsCap in the last 10 epochs. The learning rates of the 0th, 70th, and 150th epoch are 10^{-3}, 10^{-4}, and 10^{-5}, respectively. We train our SportsCap on 4 NVidia 2080Ti GPUs, and the process takes 10 hours for Motion Embedding Module, 5 hours for Action Parsing Module, and 2.5 hours for the whole SportsCap. Once trained, the network processes the $90 \times 256 \times 256$ video data at 0.5s for Motion Embedding Module, 0.05s for Action Parsing Module and 1.0s for the data fetching.

For fair comparisons, we re-train HRNet (Sun et al., 2019) and SimpleBaseline (Xiao et al., 2018) using our SMART dataset. For SCADC (Parmar and Morris, 2019b), C3D-LSTM (Parmar and Tran Morris, 2017), C3D-AVG (Parmar and Morris, 2019b), and R2+1D (Tran et al., 2018), we first pre-train the corresponding networks using the UCF101 dataset (Soomro et al., 2012) and I3D (Sun et al., 2019) on the Kinetics dataset (Sun et al., 2019), replace their output or the regression layers with our proposed Semantic Attributes

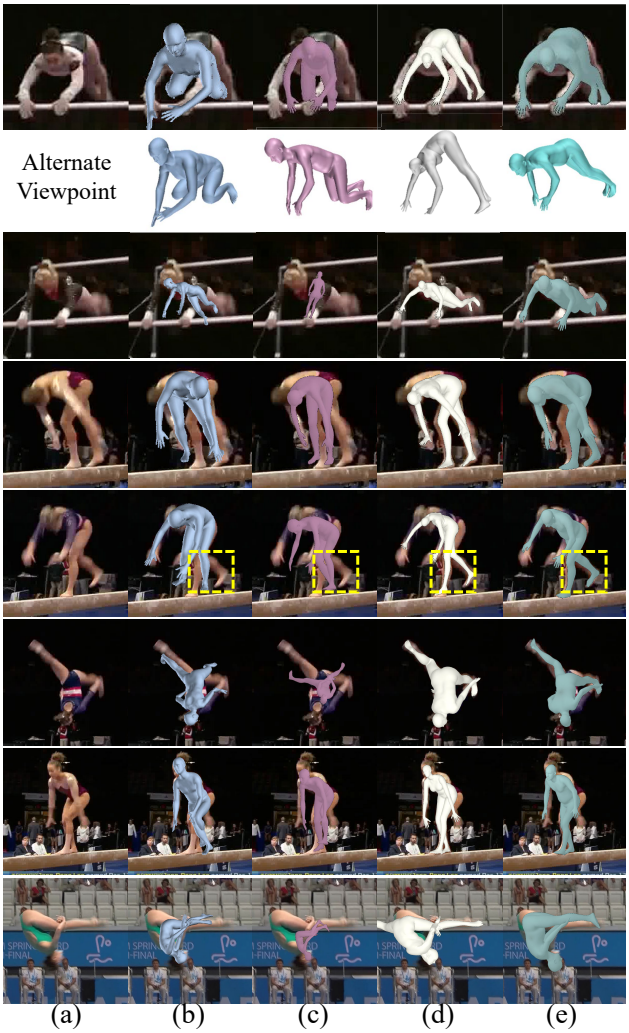


Fig. 12 Qualitative comparison with the state-of-the-art human shape recovery methods. (a) The input images. (b), (c), (d) are the results of HMR (Kanazawa et al., 2018a), VIBE (Kocabas et al., 2020), and SMPLify-X (Pavlakos et al., 2019) respectively. We fine-tune HMR and VIBE on our training set, and provide the 2D poses of our method for SMPLify-X. (e) our results of SportsCap On SMART dataset.

Mapping Block module, and fine-tune Semantic Attributes Mapping Block with our SMART dataset.

5.3 Motion Embedding Module Evaluation.

Our Motion Embedding Module aims to embed the pose motion within a certain sub-move of sports actions into parametric space. It relies on both the sub-move labels and per-frame annotated pose for evaluation, hence we test the Motion Embedding Module on the SMART dataset as few other datasets provide both information. We use the classifier in (Hu and Qi, 2019) to first predict the sub-move label. We observe this technique can achieve high accuracy, and the predicted sub-move label helps the Motion Embedding Module for pose and shape recovery. Tab. 3 shows our sub-move classification produces an average accuracy of 96.7%.

Table 3 Performance of the fine-grained sub-motion classification using the WS-DAN (Hu and Qi, 2019) on competitive diving video.

Method	Take	Twist	Some	Entry	Splash	Avg.
WS-DAN	97.8	94.5	96.5	95.3	97.6	96.7

Ablation experiments on pose estimation. There are two ablation experiments. One experiment verifies that it is more effective to use motion embedding analysis for different motion spaces by comparing PCA analysis in different motion spaces. Another experiment validates the mutual gain between 3D motion capture and fine-grained motion understanding by comparing with the state-of-the-art pose estimation works, which illustrates that mutual gain will make the 2D keypoint projection results more accurate. PCA analysis in Fig. 10 demonstrates that the poses of each sub-motion lay in a low-dimensional parametric space. As can be seen in the figure, when the number of pose coefficients is the same, if the action is restricted to each semantic pose space, the accuracy of the relative cumulative variance ratio and PCK0.3 results are higher than that under all training data pose space. This proves that motion embedding analysis for different motions is necessary and effective. So comparison with conducting PCA on the complete action data, we further reduce the space dimension by cooperating with the semantic action labels.

Tab. 2 shows the performance of the Motion Embedding Module compared with the state-of-the-art pose estimation work, including the HRNet (Sun et al., 2019) and the SimpleBaseline (Xiao et al., 2018). We also compare with state-of-the-art 3D human shape recovery methods, including the HMR (Kanazawa et al., 2018b) and VIBE (Kocabas et al., 2020). For HMR and VIBE, we project the joints of the recovered 3D human models into images and then use the projected joints as their predicted poses. For a fair comparison, we re-train the HRNet, SimpleBaseline, and et.al. on the SMART dataset. We consider the predicted joints with distance errors less than 0.3 and torso length errors less than 0.5 as the correct predictions (see details in (Andriluka et al., 2014b)) and report the percentage of correct keypoints (PCK-0.3 and PCK-0.5) as our metrics. For the classifier, we test ResNet-50, -101, -152 as our backbones and find that about every 50 more layers lead to an above 2% increase for PCK-0.3. Our approach has the similar performance as HRNet (Sun et al., 2019) and SimpleBaseline(Xiao et al., 2018) for the take-off pose but can better handle other sub-moves which exhibit more complex occlusion patterns, as shown in Fig. 11. VIBE and HMR perform better on sub-motion 0 (SM-0) which is the preliminary motion of diving while performing worse on sub-motion 1 (SM-1) which is the twist motion. Twisting and other professional motions

involve fast rotation and flipping of the body and etc, and our method benefits from the proposed sub-action motion embedding analysis which has structure constraints by PCA for similar poses in each sub-motion and has real human poses as templates, so our method can achieve higher accuracy. Our method leads by 4.3∼4.9% on PCK-0.3, and 7.1∼8.5% on PCK-0.5, this indicates our approach is more accurate in generating joint locations and the joints are more natural and stable (Fig. 11). Further, there is an approximately 5% drop in accuracy when trained the Motion Embedding Module without the prior loss provided by Sports Pose Embedding Spaces (Ours+$\mathcal{L}_{\alpha}$), demonstrates the effectiveness of the Sports Pose Embedding Spaces.

Comparison on shape recovery. Shape recovery from a single viewpoint image/video is a difficult problem, and it becomes even more challenging for sports because of severe occlusions and motion blur. While our Motion Embedding Module has the ability to accurately recover the human pose, and here we conduct the shape recovery experiments to demonstrate its performance. Tab. 2 already shows the quantitative comparison of the 2D poses, and we can see that our method outperforms the state-of-art method.

Fig. 12 qualitatively compares our method with the state-of-the-art 3D shape recovery methods. We fine-tune HMR (Kanazawa et al., 2018b) and VIBE (Kocabas et al., 2020) with our dataset, and provide our results on 2D poses for SMPLify-X (Pavlakos et al., 2019). In the figure, we show both the recovered human shape rendered at the original and alternative viewpoints. In Fig. 12, we can see that VIBE recovers more accurate 3D human shape although its 2D keypoints accuracy is lower than that of HMR (Tab. 2). For challenging sports movements, i.e. the handspring and somersault, both HMR and VIBE perform poorly while ours achieve much more accurate estimation. Further, our method has the ability to recover the limbs with higher fidelity than HMR and VIBE, as denoted by the colored box. Though the SMPLify-X method provides good 2D keypoint estimations 2, but doesn't produce as good 3D shape results. Our approach outperforms all methods by generating both accurate 2D keypoints and well-matched 3D human shapes.

5.4 Action Parsing Module Evaluation.

Next, we evaluate the Action Parsing Module module using the outputs of Motion Embedding Module. The multi-stream structure along with SAs of the Action Parsing Module module enables faster convergence. The 2s-AGCN structure (J- and B-Stream only) takes 70 epochs to converge, whereas the multi-stream structure converges after only 50 epochs. With Semantic Attributes Mapping Block, we further accelerate convergence to 10 epochs. The result shows the use of P-Stream (Tab. 4 row 6) significantly improves the accuracy vs. baseline (Tab. 4 row 4, 5), even though the accuracy on the somersault attribute drops (Tab. 4 row 5). This is expected as the network easily focuses only on one attribute without structure like semantic attributes mapping block.

We further conduct experiments to illustrate the benefits of Semantic Attributes Mapping Block over black-box. Specifically, we keep the multi-stream backbone but replace Semantic Attributes Mapping Block with black-box without using attribute loss $\mathcal{L}_{attr}$. The network with Semantic Attributes Mapping Block converges after 10 epochs with the final accuracy of 82.2% whereas the one with black-box converges after over 30 epoch with the final accuracy of 78.0%.

Comparison on action parsing. Mid-level sub-action motion embedding analysis and 3D motion capture also helps action parsing. As shown in Tab.4. We compare the overall action parsing performance of diving with methods including MSCADC (Parmar and Morris, 2019b), C3D-LSTM (Parmar and Tran Morris, 2017), C3D-AVG (Parmar and Morris, 2019b), I3D (Carreira and Zisserman, 2017) and R2+1D (Tran et al., 2018). We compare SportsCap with C3D-LSTM and R2+1D on only SMART for dive number prediction. We did not test on MSCADC, C3D-AVG, and Nibaili (Nibali et al., 2017) since they do not have the ground truth Dive Number. SportsCap achieves the highest accuracy. For action parsing, when our method using P stream information, the performance will be much better than not using it, this proves that using the pose coefficient analyzed by motion embedding benefit action parsing a lot. On the AQA dataset, we also compare SportsCap with Nibaili (Nibali et al., 2017). SMART achieves slightly better performance as C3D-AVG and outperforms MSCADC and Nabaili. This shows that our motion embedding method is not only effective on our dataset but also effective on other datasets with the same pose space. As shown in Tab. 5. On the FineGym dataset, we also do sub-action motion embedding analysis with the help of our 3D motion capture data. It can be seen from the experimental results that our method performs much better on the FineGym dataset than ST-GCN that also uses pose information. This is mainly because our 3D motion capture data and motion embedding analysis can better parse this type of sports motion under specific pose spaces.

Action Assessment. Here, we further evaluate our approach for overall detailed action assessment using the diving motion, which relies on the dive number for final motion scoring. Action assessment in diving depends on two independent scores, the difficulty score determined by the dive number, and the execution score representing the quality of the action and the splash. We reuse the extracted spatial-temporal features for action quality assessment. To evaluate the splash, we use a pre-trained C3D block (Tran et al., 2015) on the UCF101 dataset (Soomro et al., 2012) to extract a 4,096 di-

Table 4 Action parsing evaluation using state-of-the-art approaches vs. our Action Parsing Module method with joint/joint+bone/joint+bone+pose(coefficients) streams, on the SMART dataset, the FineGym dataset (Shao et al., 2020), and the AQA dataset (Parmar and Morris, 2019b).

Dataset	Method	TakeOff	ArmStand.	Twist No.	Some No.	Position	Diving No.
SMART	C3D-LSTM(Parmar and Tran Morris, 2017)	43.1	85.3	66.1	46.8	56.0	27.3
	R2+1D(Tran et al., 2018)	34.9	84.4	64.2	44.9	55.6	26.1
	I3D(Carreira and Zisserman, 2017)	61.5	92.7	70.6	69.7	70.6	58.6
	Ours+J	85.3	97.5	78.0	82.9	75.6	65.0
	Ours+J+B	84.1	98.6	76.8	**90.2**	84.7	67.1
	Ours+J+B+P Black-Box	-	-	-	-	-	78.0
	Ours+J+B+P Semantic Attributes Mapping Block	**96.4**	**99.8**	**89.5**	86.5	**92.6**	**82.2**
AQA	Nibali(Nibali et al., 2017)	74.8	98.3	78.7	77.3	79.9	-
	MSCADC(Parmar and Morris, 2019b)	78.4	97.5	84.7	76.2	82.7	-
	C3D-AVG(Parmar and Morris, 2019b)	96.3	99.7	97.5	**96.9**	93.2	-
	Ours	**97.5**	**99.8**	**97.9**	96.3	**94.0**	-

Table 5 Action parsing evaluation using state-of-the-art approaches vs. our method on FineGym dataset

Dataset	Method	VT	UB	partial Gym288
FineGym	Random	16.7	6.7	0.3
	ST-GCN	19.5	13.7	11.0
	TRN-2stream	31.4	83.0	42.9
	Ours	**34.2**	**85.7**	**46.9**

Table 6 Action assessment comparisons on SMART and AQA dataset (Parmar and Morris, 2019b), and compared with C3D-LSTM (Parmar and Tran Morris, 2017) and R2+1D(Tran et al., 2018).

Metric	Method	SMART Dataset	AQA Dataset
Sp. Cor.	C3D-LSTM	53.7	84.9
	R2+1D	55.6	**89.6**
	Ours	**61.7**	86.2

mensional feature. We concatenate this feature to our spatial-temporal feature for the final execution score assessment. Instead of treating it as a regression problem, we discretize the score range 0-100 to 49 labels evenly. We use the cross entropy loss Eq. 10 and conduct the same training strategy with the SA learning. We train our Motion Embedding Module module on our dataset for pose, while training and testing Action Parsing Module for execution score and final score on our dataset and AQA dataset respectively, Tab. 6 shows the result of the testing results. The reported execution score reaches 61.7% spearman's rank correlation (Sp. Cor.) (Pirsiavash et al., 2014).

5.5 Limitation and Discussion

As the first trial to explore the problem of joint 3D human motion capture and fine-grained motion understanding from monocular challenging non-daily video input, the proposed SportsCap still owns limitations as follows.

By tailoring our network for a specific subset of moves, our approach may generate erroneous estimations on degraded images. There are some instances such as when the image does not fall into any of the pre-defined pose categories, some body parts are invisible (e.g, head entry into the water), or incomplete images. We note that baseline algorithms in those cases may perform better as they enforce the final estimation to obey more general pose settings. Our approach, in essence, exploits the semantic action analysis, for human pose estimation. This is different from HRNet (Sun et al., 2018) or SimpleBaseline (Xiao et al., 2018) that separately predict individual joints. Consequently, our network, once trained, only tackles specific sports rather than general movements as in prior art. In addition, our approach may generate erroneous estimations for large body parts outside the viewport. Lastly, our method can tolerate common deviations from the standard movement as we purposely add such cases into the dataset. But when an athlete makes rare mistakes, it is difficult for our method to accurately detect and analyze the situation. Our current setup assumes a single video stream as input. In sports, it is common practice to show two or more video streams. In the future, we plan to combine multiple streams for a 3D pose/shape task. In addition, general human activities can always be separated into small sub-moves. Hence we plan to extend our work to general-purpose pose estimation through human action decomposition, including to expand our SMART dataset as a more general dataset. Besides, it's also an interesting direction to combine the NLP techniques to provide more natural and detailed illustrations and understanding for action assessment.

6 Conclusions and Discussion

We have presented the first approach for monocular markerless 3D motion capture and understanding for professional non-daily motions and a new dataset consisting of various challenging sports video clips with rich manually annotated 2D/3D poses and fine-grain action labels. The key insight of our approach is to utilize the semantic and temporally structured sub-motion prior in the motion embedding space and formulate the joint motion capture and understanding task in a data-driven multi-task manner. Our motion embedding module achieves robust 3D motion details reconstruction from implicit motion embedding parameters, while our novel multi-stream ST-GCN, as well as the semantic attribute mapping block, enable accurate fine-grained semantic action attributes prediction for various understanding applications like action assessment or motion scoring. Our experimental results demonstrate the effectiveness of SportsCap for both compelling 3D motion capture and fine-grained semantic action attribute reconstruction in various challenging sports scenarios, which compares favorably to the state-of-the-arts. We believe that it is a significant step to enable robust 3D motion capture and fine-grained understanding, with many potential applications in VR/AR, gaming, action recognition, and performance evaluation for gymnastics, sports, and dancing.

Acknowledgements This work was supported by NSFC programs (61 976138, 61977047), the National Key Research and Development Program (2018YFB2100500), STCSM (2015F0203-000-06) and SHMEC (2019-01-07-00-01-E00003).

References

Andriluka M, Pishchulin L, Gehler P, Schiele B (2014a) 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

Andriluka M, Pishchulin L, Gehler P, Schiele B (2014b) 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, pp 3686–3693

Anguelov D, Srinivasan P, Koller D, Thrun S, Rodgers J, Davis J (2005) Scape: shape completion and animation of people. In: ACM SIGGRAPH 2005 Papers, pp 408–416

Bertasius G, Feichtenhofer C, Tran D, Shi J, Torresani L (2018) Learning discriminative motion features through detection. arXiv preprint arXiv:181204172

Caba Heilbron F, Escorcia V, Ghanem B, Carlos Niebles J (2015) Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition, pp 961–970

Cao Z, Martinez GH, Simon T, Wei S, Sheikh YA (2019) Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence pp 1–1, DOI 10.1109/TPAMI.2019.2929257

Carreira J, Zisserman A (2017) Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 6299–6308

Choutas V, Weinzaepfel P, Revaud J, Schmid C (2018) Potion: Pose motion representation for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7024–7033

Collet A, Chuang M, Sweeney P, Gillett D, Evseev D, Calabrese D, Hoppe H, Kirk A, Sullivan S (2015) High-quality streamable free-viewpoint video. ACM Transactions on Graphics (ToG) 34(4):1–13

Defferrard M, Bresson X, Vandergheynst P (2016) Convolutional neural networks on graphs with fast localized spectral filtering. In: Advances in Neural Information Processing Systems 29, pp 3844–3852

Dou M, Khamis S, Degtyarev Y, Davidson P, Fanello SR, Kowdle A, Escolano SO, Rhemann C, Kim D, Taylor J, et al. (2016) Fusion4d: Real-time performance capture of challenging scenes. ACM Transactions on Graphics (TOG) 35(4):1–13

Fani M, Neher H, Clausi DA, Wong A, Zelek J (2017) Hockey action recognition via integrated stacked hourglass network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 29–37

Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1933–1941

Henaff M, Bruna J, LeCun Y (2015) Deep convolutional networks on graph-structured data. arXiv preprint arXiv:150605163

Hu T, Qi H (2019) See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. arXiv preprint arXiv:190109891

Hussein N, Gavves E, Smeulders AW (2019) Timeception for complex action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 254–263

Kanade T, Rander P, Narayanan P (1997) Virtualized reality: Constructing virtual worlds from real scenes. IEEE multimedia 4(1):34–47

Kanazawa A, Black MJ, Jacobs DW, Malik J (2018a) End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7122–7131

Kanazawa A, Black MJ, Jacobs DW, Malik J (2018b) End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Regognition (CVPR)

Kanojia G, Kumawat S, Raman S (2019) Attentive spatio-temporal representation learning for diving classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 0–0

Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1725–1732

Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In: Proceedings of the International Conference for Learning Representations

Kocabas M, Athanasiou N, Black MJ (2020) Vibe: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5253–5263

Li C, Cui Z, Zheng W, Xu C, Yang J (2018a) Spatio-temporal graph convolution for skeleton based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence

Li R, Wang S, Zhu F, Huang J (2018b) Adaptive graph convolutional neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence

Li Y, Li Y, Vasconcelos N (2018c) Resound: Towards action recognition without representation bias. In: Proceedings of the European

Conference on Computer Vision, pp 513–528
Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision, Springer, pp 740–755
Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) Ssd: Single shot multibox detector. In: European conference on computer vision, Springer, pp 21–37
Loper M, Mahmood N, Black MJ (2014) Mosh: Motion and shape capture from sparse markers. ACM Transactions on Graphics (TOG) 33(6):1–13
Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ (2015) SMPL: A skinned multi-person linear model. ACM Trans Graphics (Proc SIGGRAPH Asia) 34(6):248:1–248:16
Luvizon DC, Picard D, Tabia H (2018) 2d/3d pose estimation and action recognition using multitask deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5137–5146
Mahmood N, Ghorbani N, Troje NF, Pons-Moll G, Black MJ (2019) Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE International Conference on Computer Vision, pp 5442–5451
Newcombe RA, Fox D, Seitz SM (2015) Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 343–352
Nibali A, He Z, Morgan S, Greenwood D (2017) Extraction and classification of diving clips from continuous video footage. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 38–48
Pan JH, Gao J, Zheng WS (2019) Action assessment by joint relation graphs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)
Parmar P, Morris B (2019a) Action quality assessment across multiple actions. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, IEEE, pp 1468–1476
Parmar P, Morris BT (2019b) What and how well you performed? a multitask learning approach to action quality assessment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
Parmar P, Tran Morris B (2017) Learning to score olympic events. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 20–28
Pavlakos G, Choutas V, Ghorbani N, Bolkart T, Osman AAA, Tzionas D, Black MJ (2019) Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)
Pirsiavash H, Vondrick C, Torralba A (2014) Assessing the quality of actions. In: European Conference on Computer Vision, Springer, pp 556–571
Pishchulin L, Andriluka M, Schiele B (2014) Fine-grained activity recognition with holistic and pose based features. In: Proceedings of the German Conference on Pattern Recognition, Springer, pp 678–689
Pishchulin L, Insafutdinov E, Tang S, Andres B, Andriluka M, Gehler PV, Schiele B (2016) Deepcut: Joint subset partition and labeling for multi person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4929–4937
Raaj Y, Idrees H, Hidalgo G, Sheikh Y (2019) Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4620–4628
Ran L, Zhang Y, Zhang Q, Yang T (2017) Convolutional neural network-based robot navigation using uncalibrated spherical images. Sensors 17(6):1341
Shao D, Zhao Y, Dai B, Lin D (2020) Finegym: A hierarchical video dataset for fine-grained action understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2616–2625
Shi L, Zhang Y, Cheng J, Lu H (2019) Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 12026–12035
Si C, Jing Y, Wang W, Wang L, Tan T (2018) Skeleton-based action recognition with spatial reasoning and temporal stack learning. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 103–118
Soomro K, Zamir AR, Shah M (2012) A dataset of 101 human action classes from videos in the wild. Center for Research in Computer Vision
Sun K, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
Sun X, Xiao B, Wei F, Liang S, Wei Y (2018) Integral human pose regression. In: Proceedings of the European Conference on Computer Vision, pp 529–545
Tang Z, Peng X, Geng S, Wu L, Zhang S, Metaxas D (2018) Quantized densely connected u-nets for efficient landmark localization. In: Proceedings of the European Conference on Computer Vision, pp 339–354
Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp 4489–4497
Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M (2018) A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp 6450–6459
Varol G, Laptev I, Schmid C (2017) Long-term temporal convolutions for action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 40(6):1510–1517
Wen YH, Gao L, Fu H, Zhang FL, Xia S (2019) Graph cnns with motif and variable temporal block for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 33, pp 8989–8996
Xiao B, Wu H, Wei Y (2018) Simple baselines for human pose estimation and tracking. In: Proceedings of the European Conference on Computer Vision, pp 466–481
Xiaohan Nie B, Xiong C, Zhu SC (2015) Joint action recognition and pose estimation from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1293–1301
Yang Y, Ramanan D (2011) Articulated pose estimation with flexible mixtures-of-parts. In: CVPR 2011, IEEE, pp 1385–1392
Zhang W, Zhu M, Derpanis KG (2013) From actemes to action: A strongly-supervised representation for detailed action understanding. In: Proceedings of the IEEE International Conference on Computer Vision
Zhang X, Xu C, Tian X, Tao D (2019) Graph edge convolutional neural networks for skeleton-based action recognition. IEEE Transactions on Neural Networks and Learning Systems
Zhou B, Andonian A, Oliva A, Torralba A (2018) Temporal relational reasoning in videos. In: Proceedings of the European Conference on Computer Vision, pp 803–818